42p

NASA SP-15

N63-11505
Code 1 41 p

SESSION I

Celestial Mechanics and Space Flight Analysis

Chairman, CLARENCE R. GATES

Dr. Clarence R. Gates, *a native of Belvidere, Illinois, received a B.S. in Electrical Engineering from the University of Oklahoma in 1947 and a Ph. D. in Electrical Engineering and Mathematics from the California Institute of Technology in 1951. Since his graduation from Cal Tech, Dr. Gates has been associated with Jet Propulsion Laboratory, where he is currently Chief of the Systems Analysis Section. He has been active in the fields of space guidance and applied celestial mechanics and has published numerous papers in these fields.*

Introduction

By Clarence R. Gates

CELESTIAL MECHANICS: HISTORICAL BACKGROUND

For two thousand years the central problem of science was to explain and predict the motion of the planets. The solution to this problem, when it came, came swiftly; we are all familiar with the celebrated work of Brahe, Kepler, and Newton. Even so, the first steps were conservative. For example, Copernicus, who published his De Revolutionibus in 1543, held that the motion of the planets was circular, and he added epicycles in order to explain the deviations. Brahe, on the other hand, who made his celebrated observations between 1575 and 1595, rejected Copernicus and held that the earth was stationary, while Kepler, who published his well-known Laws in 1609, used magnetism to help explain the motion of the planets in elliptical paths. Galileo, who in 1610 announced the results of his observations of the Moons of Jupiter and the phases of Venus, overwhelmingly confirmed the Copernican hypothesis that the planets do in fact move about the Sun, although a few of Galileo's more fanatical opponents refused to look through the new instrument, asserting that if God had meant man to use such a contrivance in acquiring knowledge, He would have endowed man with telescopic eyes.

With the publication of Newton's Principia in 1685, an incredible achievement, the final member of the structure was beautifully fitted into place. It is interesting, here, to note that Newton's contemporaries were on the right path. For example, Robert Hooke in 1674 wrote as follows:

> . . . I shall explain a system . . . answering in all things to the common rules of mechanical motions. . . . First, that all celestial bodies whatsoever have an attraction or gravitating power toward their own centers whereby they attract not only their own parts . . . but that they do also attract all the other celestial bodies. . . . Second, that all bodies whatsoever that are put into a direct and simple motion will so continue to move forward in a straight line until they are by some other effectual powers deflected. . . . Third, that these attractive powers are so much the more powerful in operating by how much nearer the body wrought upon is to their own centers. . . .

In the next century, Euler, Lagrange, Laplace, and others provided a mathematical framework which is the foundation of modern Celestial Mechanics.

As late as the early Nineteenth Century a man of the towering stature of Gauss could find Celestial Mechanics a challenging field in which to work, but by this time the structure seemed complete; subsequently, the main stream of science and technology moved elsewhere. For those who remained to carry on the work of the past masters, the principal problems appeared to be in numerical analysis.

CELESTIAL MECHANICS AND SPACE EXPLORATION

In recent times the space program has intensified interest in Celestial Mechanics. In addition, a new field of technology, concerned with the flight path of spacecraft, has evolved. This field has variously been called astrodynamics, systems analysis, space dynamics, etc.; I shall call it Space Flight Mechanics.

Space Flight Mechanics encompasses Celestial Mechanics, since the motion of the planets must be known for the flight of a spacecraft, and since the spacecraft is acted upon by the same forces as the planets. It also encompasses geodesy, since many of our spacecraft are earth satellites, and all of our spacecraft are launched and tracked from the earth. The guidance of a spacecraft is also included in this field since, for many spacecraft, the flight path may be altered in flight by rocket motors placed in the spacecraft. Since Space Flight Mechanics requires high-speed digital computers, numerical analysis is also included.

The development of Space Flight Mechanics has been possible only through the concurrent development of modern high-speed digital computers. At the same time Celestial Mechanics, which is intimately linked with numerical analysis, is being markedly affected by digital computers; many problems are better solved by new formulations especially adapted to the computer, rather than by the machine usage of older methods. Thus, Celestial Mechanics and Space Flight Mechanics are developing simultaneously.

Some fusion of these fields is occurring. Space flight permits new knowledge to be obtained. For example, satellites of the earth, moon, or planets enable one to find the mass distribution of those bodies; and spacecraft flying to the moon or the planets can yield more exact information about their motions. Such knowledge is of importance in the cosmology of of the solar system.

Space Flight Mechanics is expanding rapidly, and qualified people are difficult to find. Most of the research and development organizations active in this field have found it necessary to train their own people. An ideal educational background would embrace elements of engineering, physics, applied mathematics, and astronomy: in engineering, control systems analysis, especially in the presence of noise, and optimization theory; in physics, classical mechanics; in applied mathematics, a solid background including numerical analysis, linear algebra, calculus of variations, and mathematical statistics; and in astronomy, Celestial Mechanics.

19. Applications of Numerical Analysis and Computer Techniques in Celestial Mechanics

By Paul R. Peabody

Dr. Paul R. Peabody, *a Research Group Supervisor, is responsible for supervision in research and development of numerical techniques at Jet Propulsion Laboratory. Dr. Peabody received his academic degrees from the University of Illinois, which conferred upon him the B.S. in Mathematics in 1948, the M.S. in Mathematics in 1949, and the Ph. D. in Mathematics in 1959. A native of Taylorville, Illinois, he is a member of ACM and Sigma Xi.*

INTRODUCTION

Every major aspect of the current space exploration program requires continual recourse to the digital computer as a tool for analysis, design, and flight operation. Sometimes this application is direct, as in an orbit determination program for an interplanetary mission. In other cases it is indirect, as in furnishing a position-velocity ephemeris of the target planet for such a mission, or for radar observations of the planet.

Application of computer techniques to such problems in celestial mechanics necessarily stimulates new researches in numerical analysis. These activities are explored in this paper by examining four major and related problems with which the computer center at the Jet Propulsion Laboratory is concerned. Time limitations preclude comprehensive reports, and detailed problem statements and presentation of results are sacrificed in order to isolate the more interesting and important topics currently being studied.

The first problem considered is that of accurate calculation of trajectories by special perturbation methods. It is best discussed from a broader point of view: as an application to celestial mechanics of one of the basic processes of numerical analysis, the numerical solution of systems of ordinary differential equations. This topic has been studied intensively over the last decade by many mathematicians, and our contributions are only a small part of the significant advances made recently.

The other three illustrations are restricted more specifically to celestial mechanics. The first two of these concern the problems of generating position-velocity ephemerides of the planets and of the Moon, and of improving the accuracy of these ephemerides using radar observations. These problems make use of high-accuracy special perturbation trajectory calculations as discussed in the first example.

The final illustration involves using the computer to aid the development of general perturbation solutions to the motion of planets, moons, and artificial satellites. This technique is still in its infancy, although some interesting results have already been obtained.

TRAJECTORY CALCULATION BY SPECIAL PERTURBATION METHODS

The principal problem in celestial mechanics is to determine the motion of a set of bodies in a force field which is essentially gravitational. Classically this problem has referred to the motion of natural bodies—planets around the Sun, moons around their primaries, and comets.

Modern space technology has added artificial satellites and lunar and interplanetary spacecraft. The general principles are the same, but there are differences significant to computation—for example, the necessity of including nongravitational accelerations such as atmospheric drag, solar radiation pressure, and thrust.

The computational problem begins after the force field has been specified and the problem reduced to the form of a system of differential equations whose solution represents the motion. There are two attacks. One is to develop an analytic representation of the solution, customarily in the form of expansions in trigonometric series with time-dependent arguments. Such methods are called general perturbation methods; they have been used to represent the motion of planets, moons, and artificial satellites. The role of the computer in such methods is discussed in the last example.

The other approach consists of a step-by-step numerical integration of the system of differential equations. Thus, the solution is represented by a tabulation of positions and velocities vs time—that is, by an ephemeris. Methods of this sort are called special perturbation methods, and they are the only ones which have been used for the accurate calculation of trajectories of comets and of lunar or planetary spacecraft.

It is necessary first to define a few terms. The special perturbation method refers to the particular system of equations used to define the motion. Thus in Cowell's method this system is merely the equations of motion themselves; in Encke's method it is for departures from a reference motion; while in Herrick's method it is for the osculating elements of the trajectory. The process by which the ephemeris is computed from the differential equations and the initial values is called the numerical integration method. Two kinds of numerical integration methods have been widely used: Runge-Kutta one-step methods, and finite difference multistep methods. Runge-Kutta methods are flexible and rather inefficient. The finite difference methods, which in turn are particular members of a larger class called linear multistep methods, are very efficient and are the ones used in practice despite certain disagreeable features. Finally, the difference between the "true" solution of the system of differential equations and the computed ephemeris is called the computational error. Note that this error does not include the effects of approximations or errors in the force field model.

All of the important problems concerning special perturbation methods are connected with either (1) *a priori* comparison between various methods and selection of particular methods for programming on the basis of accuracy vs. computing time required to solve a given class of problems, or (2) *a posteriori* analysis of the completed program in order to estimate error accumulation as a function of problem and program parameters.

Clearly a well-developed theory of the accumulation of computational error is essential to the solution of these problems. Let us examine the present state of this theory.

Theoretical investigation and experimentation indicate that error accumulation is determined by three factors—the problem itself, numerical stability, and magnitude of the local roundoff and discretization errors necessarily introduced at each step. A single local error gives rise to a disturbance which is propagated according to the character of the differential equations. In trajectory calculation by Cowell's method, the major effect is a secular displacement in time along the orbit, corresponding to a perturbation in the mean motion. Thus a disturbance increases approximately linearly with time, while other characteristics of the orbit such as the orbital plane and the *vis viva* are well preserved.

Local roundoff errors depend only on the mechanization; to decrease these, one must get a bigger computer or use extended-precision arithmetic. Local discretization errors on the other hand depend on the problem, the numerical integration method, and the integration step size h. In Cowell's or Encke's method each of the integration techniques used in practice introduces local errors which are asymptotically of the form $C_p h^{p+2} \mathbf{x}^{(p+2)}(t)$, where C_p is a coefficient dependent on the method, p is a positive integer called the order of the method, and $\mathbf{x}(t)$ is the solution vector of the differential system.

The problem of numerical instability occurs in the linear multistep methods. A method is unstable if it permits exponential growth of a disturbance; such growth cannot be tolerated in trajectory calculations. An asymptotic theory of numerical stability has been developed recently by a number of workers. In particular, G. Dahlquist has obtained an important negative result: if the order p exceeds the number of steps k in the multistep formula by more than one, the method is unstable.

Some important topics for further research can now be called out.

1. A major question is whether or not the error propagation character is affected by the numerical integration method and its parameters. It does not seem to be, so long as one uses stable methods with suitable mechanizations, in particular the device of carrying enough guard figures so that local roundoff errors are restricted to those made in the calculation of the accelerations. If this restriction is violated then roundoff errors accumulate like $n^{3/2}$, where n is the number of steps.

2. The theory of numerical stability is only an asymptotic one as the integration step h approaches zero. All linear multisteps with order greater than one become unstable for sufficiently large step size. Experience shows that the maximum step for stability, while problem-dependent, decreases as the other increases. While some people have attacked this problem recently, we still find it necessary to rely on numerical experimentation to determine the point of crossover into instability, and continued theoretical study is important.

3. Finite difference methods are of the highest order permissible for stability. However, there are certainly asymptotically stable methods of the same order for which the coefficient C_p is appreciably smaller. Thus, many people have proposed methods apparently more efficient for trajectory calculation than the finite difference methods. However,. there is evidence of a tradeoff between decreasing C_p and increasing the danger of crossover into instability. Particular cases have been studied intensively with significant results, and it is highly desirable to generalize and extend these results to a general theory.

4. Bounds on the accumulation of computational error have been obtained by a number of people, but these estimates cannot be evaluated for any trajectory problem of actual interest. I personally feel that realistic estimates can only be obtained by well-designed programs of numerical experimentation, using such techniques as extrapolation to zero step size and comparison between solutions obtained by methods of different character. There are many questions here still open.

5. Many subsidiary techniques are required in addition to the numerical integration method itself, such as starting methods, methods for changing step size, techniques for estimating local discretization error, and methods for interpolating in the computed ephemeris. Not only is it possible to develop better techniques, but common ones have not been fully investigated, particularly with respect to the additional errors they introduce.

6. Finally, let us go back to the special perturbation methods themselves. Encke's or Herrick's method, or any other method using a reference motion, involves solving differential equations for quantities which are smaller or vary more slowly than the positions as computed from Cowell's method. Thus the former methods permit a larger step size and in some cases avoidance of extraprecision arithmetic. The savings in computer time, however, is partly offset by the additional high-precision calculation of the reference motion required. In our experience Encke's method is appreciably faster than Cowell's for accuracies of about 7 figures, but may not be faster at all if accuracies of 12 or 13 figures are required. Methods based on more complicated reference motions, such as the varicentric method or one based upon the analytic solution of the Euler two-fixed-centers problem, do not appear to compete. A comprehensive study of the efficiency of various methods has not yet been made.

PLANETARY AND LUNAR POSITION-VELOCITY EPHEMERIDES

We have spent considerable time on special perturbation trajectory calculation because of its central role in celestial mechanics. Let us now rather quickly examine three other examples. The first concerns combining general and special perturbation methods in a rather interesting way.

Space exploration problems often require ephemerides of planetary and lunar positions and velocities of highest possible accuracy. Position data are available from classical general perturbation developments in which the constants of inegration or "mean elements" have been fixed by comparison with optical observations—for example, the Improved Brown Lunar Theory; Newcomb's theories of Mercury, Venus, and the Earth-Moon system; and the Hansen theory of Mars as developed by G. Clemence. The theories of the inner planets fit observations to about 10^{-6} AU, or a few hundred kilometers. However, in no case is the theory strictly consistent with the gravitational model. First, there is the unavoidable truncation of the series expansions. In addition, there were occasional manipulation errors made in deriving the coefficients in the expansions, insertion of empirical terms to adjust these errors and to compensate for phenomena unknown at the time of the theory's development (e.g., the relativistic excess motion of the perihelion of Mercury, and the irregular rotation rate of the Earth), and finally, considerable roundoff error in the published evaluations of the theories.

While these deficiencies are not too serious so far as position data are concerned, they distort velocity predictions obtained directly, either by numerical differentiation of the tabulations or by analytic differentiation of the expansions. I estimate that velocity data so derived are significant to fewer than five figures.

Special perturbation methods as discussed above yield high-accuracy position and velocity data, which are much more nearly consistent with the gravitational model over a restricted interval of integration. But here it is necessary to provide highly accurate initial position and velocity in order to start the integration. We choose these initial values so that the positions computed from the numerical integration are the best least squares fit to the classical position predictions over the arc of integration. Thus, we fit "observations" which in turn are best fits to actual observations.

This technique was used to fit the Newcomb Venus and Earth-Moon theories over a 10-year arc, from July of 1960 to July of 1970. The maximum residuals in the sense Newcomb minus Integration were only a few units in 10^{-7} AU, well within the stated accuracies of the Newcomb theories. These position-velocity ephemerides made possible the 1961 and 1962 JPL radar observations of Venus and have been used in the design and orbit determination of the present Mariner 2 project.

We are now in the process of applying this technique to all the planets and the Moon. Our main tool, currently nearing completion, is a computer program using a Cowell finite difference integration with extended precision and a special self-starting technique and containing the mechanics of least-squares orbit determination. We estimate that 10 significant figures will be maintained in the trajectory calculation over 40 orbits of a planet.

There are still many computational and research problems to be solved: new evaluations of the classical general perturbation theories to higher accuracy, development of techniques for splining consecutive fits to lunar ephemerides, writing out a relativistic system of differential equations for the motion of Mercury, and investigations into how well the numerical integration ephemeris reflects the stated mean elements of the source theory.

CORRECTIONS TO THE AU AND THE MEAN ELEMENTS OF VENUS AND EARTH-MOON FROM RADAR OBSERVATIONS OF VENUS

Let us turn to another related problem. The position-velocity ephemerides as described above do not, of course, predict the actual motion of the planet any better than the classical source theory; their important virtue is consistency with the gravitational model. Real improvement requires correction of the mean elements—that is, fitting the theory to observations. The periodicities of the inner planets are by now well established from optical observations which range over a number of centuries, but there is still considerable uncertainty in other elements.

In 1961, the Goldstone facility of JPL was successful in making doppler and range observations of Venus over a period of about 60 days centered around the conjunction of Venus, and is repeating this experiment at present. Other radar telescopes, notably the MIT instrument at Millstone, were also successful in taking range data. The major aim of the experiment, that of establishing a corrected value of the AU in kilometers (i.e., the mean Earth-Sun distance), has already been accomplished.

However, individual determinations of the AU, each one made from a single day's range data encompassing perhaps four hours of observation per day, show an increase in the value of the AU vs time over the range of the experiment. The MIT data yield the same effect, as do our doppler data in a still more obvious way. Since the value of the AU, by definition, does not change, the culprit would seem to be errors in the Venus and Earth-Moon ephemerides. In fact, this trend was reduced by about one-half by application of the corrections deduced by R. Duncombe to the mean elements of Venus and the Earth-Moon from optical data since the time of Newcomb.

We are now in the process of attempting to derive further corrections to these mean elements by combining the optical with the new radar data. While the procedure is reasonably straightforward, it may be interesting to sketch it briefly. Parameters considered only over long arcs (such as corrections to the star catalogue, secular variations in the mean elements, and periodicities) are eliminated by substitution of Duncombe's corrected values of these quantities into his normal equations. The variance in the optical observations is estimated from Duncombe's reduced equations of condition. Duncombe used a tabulation of the Newcomb theories, and our numerical integration fits to these as described above are used to calculate range and doppler residuals. Daily blocks of radar observations are reduced to yield a mean epoch, a mean observation referred to that epoch, and a weight equal to the reciprocal of the estimated variance in the observations. The mean radar observations are reduced to form weighted normal equations and combined with the weighted Duncombe normals. These final normal equations are solved using well-known methods of regression analysis to obtain (1) new corrections, (2) significance levels of these corrections, and (3) estimates of the rms residuals of observations from the corrected theory for each block of data.

The principal problem remaining is one in celestial mechanics. The value of the AU as determined by the radar data differs from the best value determined by classical techniques by considerably more than the probable error in either. This discrepancy must be explained.

COMPUTER TECHNIQUES IN GENERAL PERTURBATION THEORIES

It is obvious that the classical planetary theories are no longer adequate for reducing the extremely accurate radar observations. One cannot turn completely to special perturbation methods because of the long arcs required to cover the historical data, and it is now desirable to develop new general perturbation planetary theories of much greater accuracy. Of equal importance, the orbits of artifical satellites can in many cases be described more efficiently by general perturbation solutions. We are relatively late-comers to this field, but are now facing problems of describing the motion of artificial satellites about the Moon, Venus, or Mars, as well as making radar range and doppler observations of Earth satellites.

The bulk of the labor here consists of algebraic manipulation of long double-argument Fourier series. It is extremely desirable to short-cut this as much as possible by mechanizing the manipulations. This requires logical and algebraical operations as well as numerical operations, and it is necessary to consider the computer more generally as an information-processing device.

Consider the case of the three-body problem of Sun, disturbed planet, and disturbing planet; other problems such as the motion of a satellite around an oblate primary, are similar. The disturbing function depends on the reciprocal of the distance between the two planets. The square of this difference can be expressed rather simply as the sum of 13 terms, each a trigonometric term containing the eccentric anomalies of the two planets in their arguments, numerical coefficients being derived from values assigned to eccentricities, semimajor axes, and mutual inclination of the planets. Thus, the first problem is to develop the square root of the inverse of this quantity in a double-harmonic Fourier series. Subsequent operations require adding and multiplying such expansions, substituting one expansion into another, and differentiating and integrating the expansions.

We have made a good start on these problems in an effort to develop a computer program for generating Hansen theories of the planets. We are now turning to the generation of artificial satellite theories by the same techniques. Here, there is no difficulty in finding research topics—they proliferate!

CLOSING REMARKS

The four problems discussed above illustrate the kind of effort we are involved in. The activities of our computing center are varied and include research in numerical analysis, problem analysis support of engineers and scientists, professional programming support, and design and implementation of data processing systems. These activities genei ate many other interesting research studies—as examples, people at our center are currently carrying on extensive exploratory work in approximation theory and have done original work in such fields as spectral analysis, numerical solutions of partial differential equations, numerical solution of two-point boundary value problems, and matrix eigenvalue calculations.

However, it is the interaction between the two disciplines of celestial mechanics and numerical analysis which has primarily concerned us here. Celestial mechanics, in the era from Galileo to Poincare, fathered much of modern mathematics. Cross-fertilization between the two disciplines has already yielded significant progress in each. There is no question that there are still many important, interesting, opportunity-filled topics for further research along these lines.

20. Analysis of Satellite Orbits for Geophysical Effects

By William M. Kaula

WILLIAM M. KAULA, *Theoretical Division, NASA Goddard Space Flight Center, joined the Goddard staff in 1960. Prior to this he was Geodesist and Division Chief of Geodetic Research and Analysis of the U.S. Army Map Service. As a researcher in geophysics, his major fields of interest are figures of the earth and moon, satellite orbits, rheology of planetary interiors, the upper atmosphere, and data analysis in geophysics. He is a member of COSPAR's working group on tracking and orbit determination, and of the International Association of Geodesy's Geodetic Satellite Commission. He is a graduate of the U.S. Military Academy at West Point, and received his M.S. degree in Geodesy from Ohio State University. He is also a member of American Geophysical Union, American Astronomical Association, and Seismological Society of America.*

SUMMARY

Energy-dissipating surface forces on a close earth satellite give rise to a spectrum of orbital variations which is continuous and irregular, confined mainly to the mean anomaly, and predominant at the low frequency end (i.e., one cycle per several days). Conservation body forces on a close satellite give rise to a discrete spectrum of variations rich at the high frequency end (i.e., one or more cycles per day).

Analysis of surface force effects had produced upper-atmospheric models closely correlated in temperature variation with fluctuations in solar ultra-violet and corpuscular radiation. Further progress is dependent on solution of the problems of response of the atmosphere to solar flux and of the interaction of a satellite with its immediate environment.

Analysis of gravitational effects on close satellite orbits has produced improved determinations of north-south variations of the earth's gravitational field down to a 20° half-wave-length, and of east-west variations down to about a 50° half-wave-length. Further progress is dependent on better observed orbits of perigee height above 800 Km, and on improved statistical techniques.

INTRODUCTION

This review discusses the use of discrepancies between observation and theory as to the positions and velocities of close artificial satellites to determine geophysical properties. Such orbital variations are not only causes of "(O–C)'s" in satellite tracking; besides instrumental error, perceptible and informative causes have been the attitude of the satellite and the effects of the medium through which the signal propagates. Orbital variations, are, however, the largest cause of residuals, and a sufficiently rich source of information to more than fill a brief review. This review will also be limited mainly to analysis of variations in the exosphere, i.e., above an altitude of 500 Km.

The problem areas connected with satellite orbit analysis can be roughly defined as: (1) the instrumental problem—obtaining accurate directions or ranges or range rates; (2) the data analysis problem—determining an "observed" orbit as it varies continually from observations which are partial spatially and intermittent temporally; (3) the celestial mechanical problem—given initial conditions plus the force vector on the satellite as a function of position, velocity, and time to deduce a theoretical orbit; (4) the satellite environment interaction problem—from the physical properties of the satellite and the environment through which it travels, to deduce the force vector on the satellite and its variation in time; (5) the morphology problem—the description of the distribution of matter and energy constituting the environment; and (6) the fundamental geophysical problem—the theoretical explanation of the matter and energy distribution and their variation in time.

Satellite orbit analysis for geophysical purposes is usually defined as obtaining answers to problem area (5) by solving problem areas (2) and (3), and the emphasis of this review will be in accordance with this definition. However, as in most types of scientific investigation, progress is made by the continual interaction of theory and experiment. We shall start by briefly discussing the forces on a satellite, their expected order of magnitude, and the consequent effects on the orbit. Combining these estimates, we obtain an expected spectrum of orbital variations. Given an idea of the spectrum and a system of tracking stations we examine the statistical problem of determining the spectrum accurately and disentangling the different effects. Finally, we shall review the results obtained, the problems outstanding, and the prospects for further improvement.

FORCES ON A SATELLITE

The dominant perturbation of a close satellite orbit is that due to the flattening of the earth. The force on a close satellite of typical size, due to the main central term, is about 10^8 dynes, while the variation in this force due to the earth's flattening is about 2x10^5 dynes. The corresponding mass distribution can be visualized as a large positive center point mass and smaller negative masses, one above and one below it. It thus can be derived geometrically that the unequal pull of the masses out of the orbital plane will cause the plane to precess, expressed as the motion of the node, the point of equator crossing, referred to inertial space. Furthermore, the variation of attraction in the plane causes a motion of the axis of the ellipse, expressed as motion of perigee. As for most physical problems, an analytic solution is surer, expressing the earth's attraction as the derivative of a scalar potential with the flattening as a second degree zonal spherical harmonic J_2P_2 (sin ϕ). Transforming the latitude into orbital inclination and argument, we obtain in addition to periodic variations, the principal secular motions (ref. 1 and 2):

Perigee motion:

$$\dot{\omega}=\frac{3nJ_2}{(1-e^2)}\,2\left(\frac{a_e}{a}\right)^2\left(1-\frac{5}{4}\sin^2 i\right)+O(J_2{}^2), \tag{1}$$

Nodal motion:

$$\dot{\Omega}=-\frac{3nJ^2}{2(1-e^2)}\,2\left(\frac{a_e}{a}\right)^2\cos i+O(J_2{}^2), \tag{2}$$

where n, e, a, i are the mean motion, eccentricity, semimajor axis, and inclination of the orbit, respectively, a_e is the earth's radius, and J_2 is the ratio of the flattening to the central term: 1.0823x10^{-3}.

In addition to J_2, there are other irregularities in the gravitational field causing forces on a typical satellite of about 200 dynes. Because of the doubly attenuating effects of extrapolation to altitude and integration of acceleration to obtain position, these variations are best expressed as spherical harmonics:

$$Y_{nm}=J_{nm}P_{nm}(\sin\phi)\cos m(\lambda-\lambda_{nm}), \tag{3}$$

where ϕ is the latitude, λ is the longitude, and J_{nm} is the ratio of the harmonic to the central term: a number of $O(10^{-6})$. The more-or-less complicated spherical harmonic Y_{nm} is best remembered as a variation which changes sign $(n-m)$ times from pole-to-pole and $2m$ times in a complete circle around the equator.

The harmonics are small enough that their effects can be expressed as linear perturbations, allowing for the secular motions of node and perigee caused mainly by the flattening. For example, for the perturbation of the node by Y_{nm} we obtain an expression (ref. 3):

$$\Delta\Omega_{nm}=\sum_{p,q}\frac{\{dF_{nmp}/di\}G_{npq}\{e\}\overline{S_{nmpq}}(\omega,\ m,\ \Omega,\ \theta)}{Ma^{n+3}\sqrt{1-e^2}\sin i\{(n-2p)\dot{\omega}+(n-2p+q)\dot{M}+m(\dot{\Omega}-\dot{\theta})\}}, \tag{4}$$

in which M is the mean anomaly, ϕ is the Greenwich Sidereal Time, F_{nmp} and G_{npq} are functions of the inclination and eccentricity, respectively, and $\overline{S_{nmpq}}$ is a sine or cosine of a combination of integral multiples of its stated arguments. Considering that G_{npq} is $O(e|^{q}|)$ and that the orders of magnitude of the secular terms in the denominator are:

$$O(\dot{M})=10 \text{ cycles/day},$$

$$O(\dot{\theta})=1 \text{ cycle/day},$$

$$O(\dot{\omega})=0.01 \text{ cycle/day},$$

$$O(\dot{\Omega})=0.01 \text{ cycle/day},$$

we see that the dominant term for a particular harmonic will be for the subscript combination $n-2p+q=0$, $q=-1$, 0, or 1, and that terms of small m will have larger effects than those of large m—in particular, the zonal harmonics, for which $m=0$.

Turning now to surface forces, if we treat the momentum interchange of a satellite of velocity V with the air molecules through which it moves as a purely mechanical problem and assume the air molecules to have a mean free path appreciably larger than the satellite and velocity appreciably smaller, we obtain a force (ref. 4):

$$F_d=\frac{C_D}{2}A\rho V^2, \qquad (5)$$

where C_D, about 2.2, depends on the shape of the satellite and the manner of reflection of the air molecules; A is the cross sectional area; and ρ is the air density. Taking a typical satellite at an altitude of 500 km, we find the drag force is less than 10 dynes—i.e., about two orders of magnitude less than the forces due to the irregularities of the gravitational field. However, the drag force vector is always directed contrary to the velocity vector, and hence is not "averaged out" by the rotation of the earth and the revolution of the satellite, as are the gravitational effects. The resulting energy loss causes a contraction of the orbit and a speeding up of the satellite to counteract the increased gravitational pull. If we further consider the rapid decrease of density with altitude, the drag on an eccentric orbit can be considered virtually as an impulse at perigee. Combining this energy loss with the energy equation,

$$\Delta V^2 \approx \Delta V_p^2=\mu\Delta\left[\frac{2}{r_a}-\frac{1}{a}\right]=\mu\Delta\left[\frac{r_p}{r_a}\cdot\frac{1}{a}\right], \qquad (6)$$

we see that since $\Delta V^2<0$, and $\Delta a<0$, necessarily $\Delta\frac{r_a}{r_p}<0$, i.e., the orbit will decrease in eccentricity.

This concentration of drag at perigee, coupled with the motions of perigee and node with respect to the sun, causes a satellite to sample different parts of the atmosphere and thus yield evidence as to its variation in space as well as time.

At altitudes in the exosphere—i.e., above about 550 km altitude, where the mean free path exceeds the scale height, the range over which the pressure drops by a factor of $1/e$—the purely mechanical model of drag is insufficient, because of the appreciable population of charged particles. The dominant property is the high velocity of the electrons as compared to the satellite velocity. These electrons impinging on the satellite cause it to acquire a negative charge. The negative charge in turn causes the satellite to acquire a sheath of positive charge, which in turn increases the drag due to both the mechanical interaction of this cloud with the air and the coulomb repulsion of the ions. This situation is distorted by the nonuniform distribution of electrons due to the magnetic field, along the lines of which the electrons will gyrate, and the impact of photons from the sun which cause ejection of electrons. The resulting nonuniform distribution of charge causes currents to flow and the intersection of the magnetic field lines by the satellite will add an inductive effect contributing to drag by tumbling of the satellite (ref. 4–6, 38).

Under these circumstances, a proper solution of the problem must consider the hydrodynamic and electromagnetic aspects together. The applicable equation is the collision-free Boltzmann transport equation (ref. 7):

$$\frac{\partial f_e}{\partial t}+c\cdot\Delta f_e+\frac{e}{M_e}\Delta\phi\Delta_c f_e=0 \qquad (7)$$

where f_e is the energy distribution of electrons, c is the velocity of the electron relative to the satellite, e and M_e are the charge and mass of an electron, ϕ is the electrostatic potential, and Δ and Δ_c are gradients with respect to position and velocity space, respectively. Also applicable are the Boltzmann equation for the ions, the Poisson equation for the potential ϕ, and the assumed unperturbed energy distributions. The critical quantity in this problem is the Debye shielding distance:

$$\lambda_D=\sqrt{\frac{kT}{4\pi ne^2}} \qquad (8)$$

where k is the Boltzmann constant, T is the temperature, and n is the number density. The Debye length is the maximum distance at which charged particles will interact significantly; in the exosphere, it is on the order of one centimeter. Solutions which have been made so far for the charged drag problem have either been numerical calculations assuming steady flow of particles (ref. 8), or analytical developments assuming the object to be smaller than a Debye length, thus permitting linearization for solution (ref. 7, 9, 10). These analytical solutions indicate that the satellite will cause oscillations in the plasma density; i. e., there will be a hydromagnetic shock wave behind the satellite, and energy will be transferred from the satellite to the medium by this "wave" drag. Further properties which are still unsure are the unperturbed energy distribution of the electrons, the extent of thermal equilibrium, and the nature of reflection of ions from a charged satellite. These uncertainties as to the interaction of a satellite with its environment indicate that "densities" in the upper exosphere derived from orbits under the assumption of neutral drag, after equation (5), must be treated with caution (ref. 6, 11, 38).

There have not been many extensive analytical developments of the dynamical effects of drag on the orbit comparable to those for gravitational perturbations, because of uncertainties as to the satellite-environment interaction; the large and irregular variations of the atmosphere; and, as indicated by the developments for even the simplest atmospheric models (ref. 12, 13), the mathematical complexity of the problem. Hence most studies of orbits for drag have used a combination of relatively simple analytical developments with numerical methods.

Treatment of the significant extraterrestrial effects on satellite orbits, the luni-solar gravitational perturbations and radiation pressure, is largely a mathematical problem, since the physics involved is better known. For drag analyses radiation pressure effects must be taken into account, since the effect of the earth's shadow is to cause energy variations comparable to those caused by drag above 500 to 1,000 km altitude, depending on the phase of the 11-year cycle of solar activity.

ANALYSIS OF SATELLITE OBSERVATIONS

Remembering that:

(1) the gravitational perturbations are relatively large (200 dynes) but oscillating in direction, and limited in frequency pattern to integral multiples of the earth's rotation rate and the rates of the various orbital angles,

(2) the drag perturbations are small (less than 10 dynes), always acting counter to the velocity vector relative to the surrounding medium, and irregularly variable with solar activity and atmospheric response thereto, we can draw a schematic diagram of the spectrum of expected variations from a secularly changing Keplerian ellipse. The spectrum shown in figure 20-1 is characteristic of about 3 months' record of an orbit of moderate eccentricity, with a perigee of around 500 km altitude. Since the

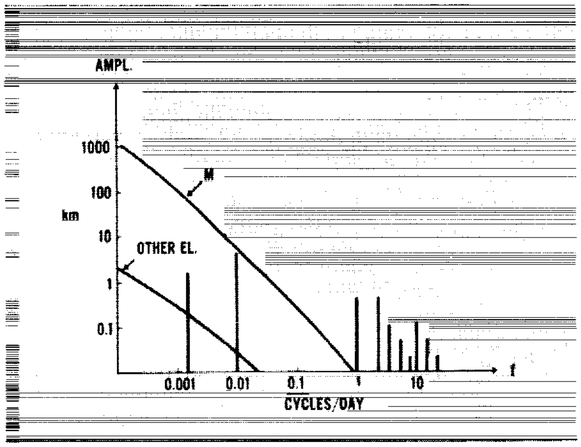


FIGURE 20-1.—Spectrum of Satellite orbit variations.

satellite state is a six-component vector rather than a scalar, figure 20–1 is a simplification of the actual situation.

The gravitational effects have a line spectrum, with a few large terms of low frequency on the order of 1 cycle/50 days, dependent on the rotation of the perigee. There will be a cluster of lines of smaller magnitude near one, two, etc. cycles/day, corresponding to integral multiples of the earth's rotation rate less the nodal motion rate, plus various multiples of the perigee rotation. Finally there will be a cluster of still smaller terms of frequency more than 10 cycles/day, dependent on the rate of revolution of the satellite itself about the earth.

The drag effects, on the other hand, have a continuous spectrum, which slopes downward very steeply toward the high frequency end, and which is some order of magnitude larger for the mean anomaly than it is for the other elements of the orbit.

If a satellite orbit and altitude specifications could be closely controlled and their variations continuously and completely observed, like a laboratory experiment, the problems of analyzing drag effects and gravitational effects would be quite distinct. For a given spherical harmonic term in the gravitational field, only certain lines could appear, and the relative magnitude of lines for different frequencies and different orbital elements would be fixed, leaving only the amplitude and phase angle to be determined. After subtracting out these gravitational effects, the remaining residuals could be subjected to cross-correlation analysis with other indicators of solar and atmospheric variation. In practice, however, we are forced to deal with satellites which have varying cross-sections, which have orbits giving an extremely biased and non-uniform sample of the atmosphere, and which are observed infrequently and incompletely by stations of non-uniform geographic distribution. This nonuniform distribution is particularly troublesome in determining gravitational variations which are functions of longitude, since a given station can observe an orbit only when the angle (GST-node—integral multiples of which are phase angles of the gravitational effects—is near one of two values corresponding to the station zenith. The limitation will also cause an error in station position to give rise to a spurious spectrum of orbital variations involving multiples of (GST-node).

For analyzing low frequency variations of one cycle per few days or more, the observations are frequent enough that we can be fairly sure that empirically determined variations of the Keplerian elements reflect mainly true variations of the orbit. Such empirically determined elements are used by Jacchia, Priester, and others (ref. 14–16) to determine slowly varying conditions of the atmosphere, and by Kozai (ref. 17) and others to determine zonal harmonics of the gravitational field. For the more high frequency variations due to drag, the fact that they affect the mean anomaly much more than any other element can be used: it can be assumed that observational residuals with respect to mean orbital elements determined over several days are due entirely to variations in the mean anomaly, and the analysis can be applied to the residuals as transformed into mean anomaly variations; this technique has been most extensively applied by Jacchia (ref. 14). For the high frequency perturbations due to longitudinal variations of the gravitational field, however, the analysis must be applied to the observational residuals themselves or else confined to orbital segments with a large number of observations. In seeking these small high frequency effects, we find the orbital characteristics vary slowly enough that the perturbations can be treated as stationary time series, and linear regression methods applied. Since such methods entail computational manipulation of arrays comparable in dimension to the number of observations, thus far there have been applied only approximate methods assuming randomness of errors between observations, and utilizing devices such as pre-weighting of parameters and low-weighting of along-track residuals compared to across-track residuals (ref. 18 and 19).

RESULTS OBTAINED FOR ATMOSPHERIC VARIATIONS

The earliest satellite orbits had acceleration rates which indicated that an appreciable increase was required in the densities of upper atmospheric models for the zone from 150 to 700 km altitude. As the perigees of these satellites relative to the sun moved, and as new satellites were put in orbit, it was found by Jacchia, Priester, King-Hele, and others that there was an appreciable bulge in atmospheric density in the general direction of the sun, but having a lag on the order of two hours behind the sun. Figure 20–2 is a representation of such a model based on satellite orbits by Martin and associates (ref. 15). To emphasize the contours of the upper atmosphere, the solid earth and the first 300 km of atmosphere have been shrunk to a point in Figure 20–2. Note that the lowest density occurs shortly before dawn; that there is a rapid rise in density until mid-afternoon, after which there is a slower decrease through the night; and that there is also an appreciable increase in density scale height—the interval over which it decreases by a factor 1/e—with altitude on the day side of the orbit, but a much smaller one on the night side.

In addition to the diurnal variation, there have been determined other correlations of satellite accelerations with indicators of solar activity, such as radiation in the 3–30 cm wave band and the A_p index of geomagnetic activity. Figure 20–3, due to Bryant (ref. 20), gives a very prominent example of such correlation: the increases in acceleration of the large balloon satellite Echo I at the time of the great solar flares of November-December 1960. Correlations with these flares have been found in the orbits of seven satellites by Jacchia (ref. 21). Such strong correlation with a short term phenomenon is exceptional, however; normally, the correlation of day-to-day accelerations with the A_p is very slightly positive, and moderate with the 10.7 cm radiation. The correlation is more marked with the 27-day variation in the 10.7 cm radiation—when it is prominent—still stronger with the semiannual variation in geomagnetic activity, and strongest of all with the 11-year cycle in solar activity. The semiannual variation in upper atmospheric density, originally found by Paetzold (ref. 22), is, in fact, more pronounced than the corresponding variation in geomagnetic activity.

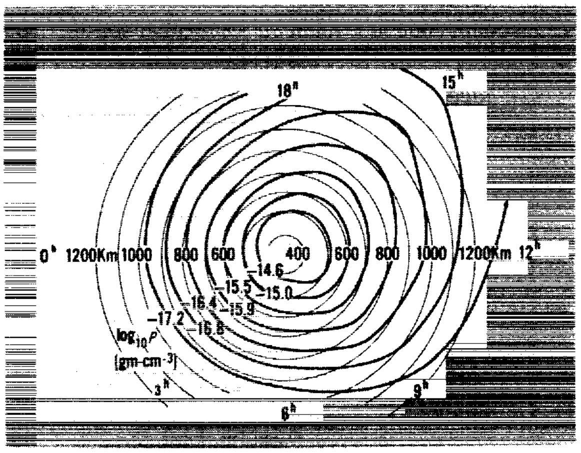


FIGURE 20–2.—Upper atmosphere from satellites—lines of equal density at scale height interval.

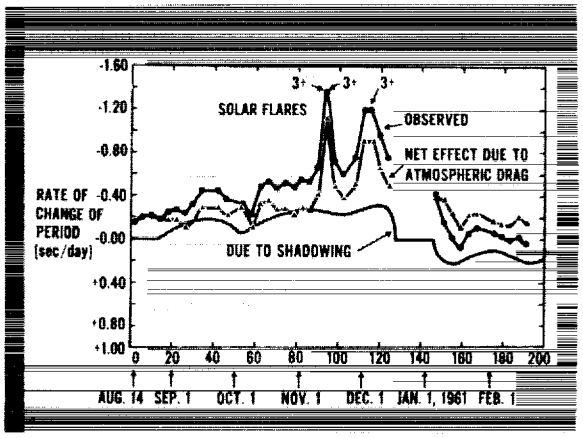


FIGURE 20–3.—Rate of change of the period of Echo I.

Because of the nonuniform chemical content of the upper atmosphere, it is now recognized that it is impossible to construct a physically consistent model of the atmosphere in terms of density variations deduced from satellite drag and the indicators of solar activity with which it is correlated. The more fundamental quantity is temperature. Theoretical consideration of photoionization rates, particle energies, and energy transfer between particles lead to the conclusion that there will be thermal equilibrium between all components of the upper atmosphere except in the ionosphere below about 400 km. This conclusion is fairly well substantiated by measurements of charged particle energies (ref. 23, 39). In addition to thermal equilibrium, there must be diffusive equilibrium because of the rarity of the upper atmosphere, i.e., each of its chemical constituents behaves independently of the others. To each constituent there is independently applicable:

(1) the equation of hydrostatic equilibrium, relating pressure, p, to density, ρ, as a function of altitude, h:

$$\frac{dp}{dh}=-\rho(h)g(h) \qquad (9)$$

where g is the acceleration due to gravity; and

(2) the equation of state for a perfect gas, relating pressure and density to molecular mass, m, and absolute temperature, T:

$$pm=\rho kT \qquad (10)$$

where k is the Boltzmann constant. Differentiating equation (10) with respect to h, substituting in equation (9), and adding together the equations for the different components yields a differential equation for the change in density as a function of altitude:

$$d\rho=-\frac{g}{kT}\left\{\sum_i \rho_i m_i\right\}dh=-\sum_i\frac{\rho_i}{H_i}dh, \qquad (11)$$

where the scale height H_i for component i is

$$H_i=\frac{kT}{m_i g} \qquad (12)$$

Hence the scale height indicated by the spacing of the lines in figure 20–2 is but a crude average of the scale heights of the individual components. The inverse proportionality of H_i to m_i further indicates that there should be a marked change from heavier to lighter constituents with altitude (ref. 24 and 25).

To analyse density variations deduced from satellite orbits the first step is thus to find a compatible number density of different chemical components; the most marked revision of this sort of atmospheric models which has been required is an appreciable increase in the proportion of helium, first suggested by Nicolet (ref. 26) to account for the higher densities above 1000 km. The next step is then to translate the densities into temperature, and thus to translate the variations in density to variations in temperature. Recent empirical models of the atmosphere, such as those of Jacchia (ref. 14) and Paetzold (ref. 27) express the correlation of upper atmosphere variation with solar activity in terms of a correlation coefficient between the temperature in degrees Kelvin and the 10.7 cm flux in 10^{-22} watts/cm^2/cycle/sec. This coefficient is about 4.5 for the long term variations associated with the 11-year cycle, but only about 2.5 for the erratic "27-day" oscillations. The correlation of temperature with the geomagnetic index A_p, in units of 2γ (2×10^{-5} gauss), is less: 1.0 to 1.5. These differences in correlation suggest that about two-thirds of the heating of the atmosphere comes from the extreme ultraviolet radiation, and one-third from corpuscular or other radiation.

Besides the described variations, there is the diurnal bulge which amounts to about 35–40 percent when translated from density into temperature. Harris and Priester (ref. 28), assuming an ultraviolet heat source below 120 km altitude and heat transport by conduction and mass flow, find that the energy input required for a variation of this magnitude is about 2.0 erg/cm^2/sec, which implies extremely high efficiency of conversion when compared to the EUV flux observed by Hinteregger (ref. 29). Furthermore, the peak density of the model is attained about 3 hours later in the day than that derived from satellite orbits. Assuming that about one-third of the heating has another source, with a peak at about 0900 local time, yields a much better fit to observations up to about 1000 km. Above 1000 km, the diurnal variations of the model are smaller than those observed.

A plausible model for heating by corpuscular radiation through hydromagnetic waves, in which oscillating charged particles collide with neutral atoms, was originally suggested by Dessler (ref. 30). However, calculations based on this model indicate that an energy dissipation on the order of 1.5 erg/cm^2/sec requires a hydromagnetic wave of 400 γ magnetic field amplitude (ref. 31). The usual amplitude of irregular fluctuations in the field at sea level is only 20 γ. Some attenuation takes place between the ionosphere and the ground, but it is not known how it could be so much. Hence it is still in doubt whether the additional heating is corpuscular or an overlooked type of photoionization.

This part of the review has ranged somewhat far from strictly an analysis of satellite orbits because, as previously mentioned, the rather awkwardly biased sample which an orbit constitutes requires a fairly good model of the environmental morphology to analyze it effectively. Much more detailed analyses of satellite accelerations could be made, but for such analyses to be useful there are needed guides for what to seek in the form of better models of the upper atmosphere: the spectra of energy inputs and transfers (see, e.g., ref. 32) and the associated energy dissipations which give rise to variations in temperature. Also, above 1000 km, there are needed better solutions of the hydromagnetic problem of the interaction of the satellite with its environment. Still more fundamental, of course, is an explanation of variations in the original source of the energy for the atmospheric variations: the sun (see, e.g., ref. 33).

RESULTS OBTAINED FOR GRAVITATIONAL VARIATIONS

Discussion of satellite orbit analysis for variations in the earth's gravitational field is appreciably simpler, because, as shown by figure 20–1, we are analyzing for a fixed line spectrum; because the morphology of the field is essentially two-dimensional, the variation with altitude being fixed by Laplace's equation; and because there are no doubts as to the nature of the satellite-environment interaction. Hence we can discuss determinations of parameters based on oscillations in the orbit on the order of 30 meters, while the smallest oscillations which have been used in connection with drag determinations are more than a kilometer.

As indicated by figure 20–1, there are a few long period variations in the orbit, corresponding to the cases when n is odd, $n-2p+q=0$, $m=0$ in equation (4). In addition, there are secular effects similar to equations (1) and (2) for the cases when n is even, $n-2p+q=0$, $m=0$. In the first two or three years of close satellite orbits, there were several analyses for the corresponding terms in the gravitational field, the zonal harmonics, which reflect purely north-south variation, e.g., those by O'Keefe and by King-Hele (ref. 34, 35). As a greater variety of orbits were obtained and observational accuracy improved, activity in this area decreased because of the increase in computation required to get appreciably improved values. In the past year, the only significant new results are those of Y. Kozai (ref. 17) based on Baker-Nunn camera observations of 13 satellites. His latest values are:

$$
\begin{array}{ll}
J_2=1082.48\times10^{-6} & J_5=-0.064\times10^{-6} \\
\quad\pm0.04, & \quad\pm0.007, \\
J_4=-1.84\times10^{-6} & J_7=-0.470\times10^{-6} \\
\quad\pm0.09, & \quad\pm0.010, \\
J_6=0.39\times10^{-6} & J_8=-0.02\times10^{-6} \\
\quad\pm0.09, & \quad\pm0.07, \\
J_3=-2.562\times10^{-6} & J_9=0.117\times10^{-6} \\
\quad\pm0.007, & \quad\pm0.011
\end{array}
$$

The uncertainties of the odd-degree zonal harmonics are smaller because the periodic variations are less subject to distortion by drag, etc. than are the secular changes. The above coefficients reflect all significant variation in a purely north-south direction of half wave length more than 20°, or about 1400 miles. The geophysical interest in these results in the sharp drop in magnitude above J_4, which suggests that the corresponding density irregularities must be rather deep in the earth's mantle.

Currently attention is directed more toward determination of the tesseral harmonics, which express variation of the gravitational field on a longitudinal as well as a latitudinal direction. As previously mentioned, the principal difficulties in determining these variations from daily, semidaily, etc. oscillations in the orbit are the nonuniform distribution of observations and the existence of errors in station positions. The nonuniformity of observation distribution is enhanced by dependence on solar illumination for the most accurate observations available, those by the Baker-Nunn cameras of the Smithsonian Astrophysical Observatory. This difficulty limits suitable satellite orbits to those of perigee height between 800 and 1500 km, and of moderate eccentricity, i.e., high enough to have moderate drag effects, and to be observed fairly often, but low enough to be perceptibly perturbed by the gravitational variations.

Results obtained by the principal investigators in this area, Kozai (ref. 17), Newton (ref.

36), and Kaula (ref. 19) vary appreciably from each other, probably as much because of the differences in statistical treatment as in the orbits employed. Their most recently published values for the lowest degree tesseral harmonic, J_{22}, (corresponding to an equatorial ellipticity $(a_1-a_2)/a_1$ six times as great) are:

Y. Kozai (ref. 17): $J_{22}=2.31\times10^{-6}$ $\lambda_{22}=37.5°W$

R. Newton (ref. 36): $J_{22}=2.15\times10^{-6}$ $\lambda_{22}=11°W$

W. Kaula (ref. 19): $J_{22}=1.62\times10^{-6}$ $\lambda_{22}=21.5°W$

Determinations have also been made of other tesseral harmonics, the most extensive analysis probably being that by Kaula (ref. 19), in which solutions were made from Baker-Nunn camera observations of satellite 1960 Zeta 2 for 28 tesseral harmonic coefficients with indices ranging up to $n, m=6, 4$, and for the 18 parameters expressing the positions of 6 geodetic datums. The unknowns selected were those expected to cause variations of ±20 meters or more, as indicated by statistics based on terrestrial geodetic data. The results for the gravitational variations are shown in figure 20–4. Individual coefficients for which a good degree of internal consistency was obtained over 300 days of orbit were J_{31}, J_{22}, J_{41}, and J_{43}. In addition, a vanishingly small value was obtained for J_{21}, as independently predicted by latitude variation observations.

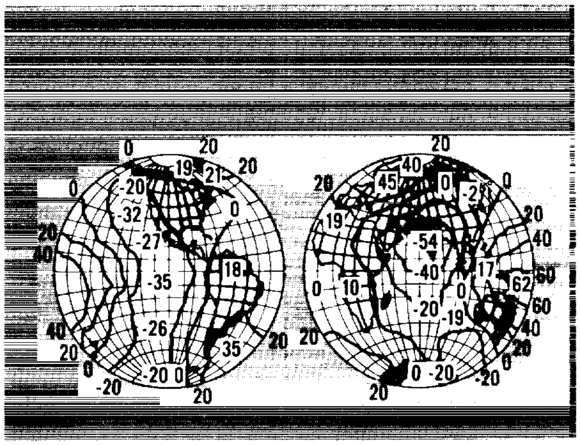


FIGURE 20–4.—Geoid Heights in meters—referred to an ellipsoid of flattening 1/298.24.

All analyses made thus far for gravitational variations have assumed that the errors of observation are random with respect to each other. Since "error" encompasses all discrepancies between observation and mathematical model, and since the model is patently inadequate at present to take into account drag variations, the assumption will produce distorted results from nonuniformly distributed observations. With the large, high-speed computers now available, and the dynamical and geometrical spadework for these effects accomplished, more rigorous linear regression techniques can be applied (ref. 18).

An even more significant improvement can be expected with improvements in instrumentation. The geodetic satellite Anna (ref. 37) is the first satellite to carry a flashing light, thus obtaining greatly improved distribution of camera observations. Eventually, radio tracking methods will become more important as they improve in accuracy, because of their ability to obtain superior distribution of observations by tracking in daylight and through clouds.

REFERENCES

1. KOZAI, Y.: *The Motion of a Close Earth Satellite.* Astronomical Jour., vol. 64, no. 9, 1959, p. 367.
2. BROUWER, D.: *Solution of the Problem of Artificial Satellite Theory Without Drag.* Astronomical Jour., vol. 64, no. 4, 1959, p. 378.
3. KAULA, W. M.: *Analysis of Gravitational and Geometric Aspects of Geodetic Utilization of Satellites.* Geophysical Jour., vol. 5, no. 2, 1961, p. 104.
4. JASTROW, R., and PEARSE, C. A.: *Atmospheric Drag on the Satellite.* Jour. Geophysical Res., vol. 62, no. 3, 1957, p. 413.
5. BEARD, D. B., and JOHNSON, F. S.: *Charge and Magnetic Field Interaction with Satellites.* Jour. Geophysical Res., vol. 65, no. 1, 1960, p. 1.
6. SHAPIRO, I. I.: *The Prediction of Satellite Orbits.* Proc. Symposium on the Dynamics of Satellites, Int. Union Th. and Appl. Mechanics, Berlin, Springer-Verlag, 1962, in press.

7. Kraus, L., and Watson, K. M.: *Plasma Motions Induced by Satellites in the Ionosphere.* Physics of Fluids, vol. 1, no. 6, 1958, p. 480.
8. Davis, A. H., and Harris, I.: *Interaction of a Charged Satellite with the Ionosphere.* NASA Tech. Note D–704, 1961.
9. Yoshiwara, H.: *Motion of Thin Bodies in a Highly Rarefield Plasma.* Physics of Fluids, vol. 4, no. 1, 1961, p. 100.
10. Greifinger, P. S.: *Induced Oscillations in a Rarefield Plasma in a Magnetic Field.* Physics of Fluids, vol. 4, no. 1, 1961, p. 104.
11. Chopra, K. P.: *Interactions of Rapidly Moving Bodies in Terrestrial Atmosphere.* Reviews of Modern Physics, vol. 33, no. 2, 1961, p. 153.
12. Brouwer, D., and Hori, G. I.: *Theoretical Evaluation of Atmospheric Drag Effects in the Motion of an Artificial Satellite.* Astronomical Jour., vol. 66, no. 5, 1961, p. 193.
13. Cook, G. E., King-Hele, D. G., and Walker, D. M. C.: *The Contraction of Satellite Orbits Under the Influence of Air Drag.* Proceedings of the Royal Society, Series A, vol. 257, 1960, p. 224; vol. 264, 1961, p. 88; vol. 267, 1962, p. 541.
14. Jaccia, L. G., and Slowey, J.: *Accurate Drag Determination for Eight Artificial Satellites; Atmospheric Densities and Temperatures.* Smithsonian Inst. Astrophysical Obs. Spec. Rep. no. 100, 1962.
15. Martin, H. A., Neveling, W., Priester, W., and Roemer, M.: *Model of the Upper Atmosphere from 130 Through 1600 Km. Derived From Satellite Orbits.* Space Research, vol. 2, Amsterdam North Holland Publ. Co., 1961, p. 902.
16. King-Hele, D. G., and Walker, D. M. C.: *Upper Atmosphere Density During the Years 1957 to 1961, Determined From Satellite Orbits.* Space Research, vol. 2, Amsterdam, North Holland Publ. Co., 1961, p. 918.
17. Kozai, Y.: *Numerical Results From Orbits.* Smithsonian Inst. Astrophysical Obs. Spec. Rep. no. 101, 1962.
18. Kaula, W. M.: *Satellite Orbit Analyses for Geodetic Purposes.* Proc. Symposium on the Dynamics of Satellites, Int. Union Th. and Appl. Mechanics, Berlin, Springer-Verlag, 1962, p. 187.
19. Kaula, W. M.: *Tesseral Harmonics of the Gravitational Field and Geodetic Datum Shifts From Camera Observations of Satellites.* Jour. Geophysical Res., vol. 67, 1962, in press.
20. Bryant, R. W.: *A Comparison of Theory and Observation of The Echo I Satellite.* Jour. Geophysical Res., vol. 66, no. 11, 1961, p. 3066.
21. Jacchia, L. G.: *The Atmospheric Drag of Artificial Satellites During the October 1960 and November 1960 Events.* Smithsonian Inst. Astrophysical Obs. Spec. Rep. no. 62, 1961.
22. Paetzold, H. K., and Zschorner, H.: *An Annual and Semiannual Variation of the Upper Air Density.* Geofisica Purae Applicata, vol. 48, 1961, p. 85.
23. Bauer, S. J., and Bourdeau, R. E.: *Upper Atmosphere Temperatures Derived From Charged Particle Observations.* Jour. Atmospheric Sciences, vol. 19, no. 3, 1962, p. 218.
24. Johnson, F. S.: *Atmospheric Structure.* Astronautics, vol. 7, no. 8, 1962, p. 54.
25. Nicolet, M.: *Density of the Heterosphere Related to Temperature.* Smithsonian Inst. Astrophysical Obs. Spec. Rep. no. 75, 1961.
26. Nicolet, M.: *Helium, An Important Constituent in the Lower Exosphere.* Jour. Geophysical Res., vol. 66, no. 7, 1961, p. 2263.
27. Paetzold, H. K.: *Corpuscular Heating of the Upper Atmosphere.* Jour. Geophysical Res., vol. 67, no. 7, 1962, p. 2741.
28. Harris, I., and Priester, W.: *Time-Dependent Structure of the Upper Atmosphere.* Jour. Atmospheric Sciences, vol. 19, no. 4, 1962, p. 286.
29. Hinteregger, H. E.: *Preliminary Data on Solar Extreme Ultraviolet Radiation in the Upper Atmosphere.* Jour. Geophysical Res., vol. 66, no. 8, 1961, p. 2367.
30. Dessler, A. J.: *Ionospheric Heating by Hydromagnetic Waves.* Jour. Geophysical Res., vol. 64, no. 4, 1959, p. 397.
31. Francis, W. E., and Karplus, R.: *Hydromagnetic Waves in the Ionosphere.* Jour. Geophysical Res., vol. 65, no. 11, 1960, p. 3593.
32. MacDonald, G. J. F.: *Spectrum of Hydromagnetic Waves in the Exosphere.* Jour. Geophysical Res., vol. 66, no. 11, 1961, p. 3639.

33. PRIESTER, W., and CATTANI, D.: *On the Semiannual Variation of Geomagnetic Activity and its Relation to the Solar Corpuscular Radiation.* Jour. Atmospheric Sciences, vol. 19, no. 2, 1962, p. 121.

34. O'KEEFE, J. A., ECKELS, A., and SQUIRES, R. K.: *The Gravitational Field of the Earth.* Astronomical Jour., vol. 64, no. 7, 1959, p. 245.

35. KING-HELE, D. G.: *The Earth's Gravitational Potential, Deduced From the Orbits of Artificial Satellites.* Geophysical Jour., vol. 4, no. 1, p. 3, 1961.

36. NEWTON, R. R.: Paper presented at the Symposium on the use of Artificial Satellites for Geodesy, Com. Space Res. and Int. Assoc. Geodesy, Washington, 1962, in press.

37. MACOMBER, M.: *Project ANNA.* Symposium on the use of artificial satellites for Geodesy, Com. Space Res. and Int. Assoc. Geodesy, Washington, 1962, in press.

38. WOOD, G. P.: *Electric Drag on Satellites.* NASA University Conf. on Science and Technology of Space Exploration, Chicago, 1962, in press.

39. BOURDEAU, R. E.: *Ionospheric Research in Space.* NASA University Conf. on Science and Technology of Space Exploration, Chicago, 1962, in press.

21. Applications of Celestial Mechanics to Spacecraft Flight

By Thomas W. Hamilton

THOMAS W. HAMILTON, *Assistant Section Chief of the Systems Analysis Section of Jet Propulsion Laboratory, was born in Evergreen Park, Illinois. He was graduated from the California Institute of Technology with a B.S. in Physics in 1952. Mr. Hamilton is a member of Sigma Xi.*

INTRODUCTION

How is celestial mechanics used in flight to the Moon and beyond? What new information will come from spacecraft flight to enrich celestial mechanics?

Table 21–I presents an outline of this paper. Three distinct subjects are listed to give an idea of the range of problems of current interest as well as to suggest the benefits of spacecraft flight to celestial mechanics.

TABLE 21–I.

Space systems analysts, of course, have different problems from those of astronomers, who have developed celestial mechanics to meet their needs. This is because space systems analysts can, to a great extent, *select* and *control* the flight path of the spacecraft from parking orbit to the mission's end. Consequently, it is necessary to investigate and describe systematically a great many possible trajectory choices (Ref. 1). The objective is to select the path which will best accomplish the mission within the constraints imposed by payload weight, scientific experiment requirements, and even the very systems which control the flight path. An interesting aspect of a Mars orbiter mission has been chosen to show some of these interactions.

It is obvious that the measurements made by the spacecraft of its surroundings should yield fruitful new information. Less obvious are the benefits of making navigational measurements from the ground. The precise two-way doppler measurements taken on the current flight of Mariner 2 to Venus can give us an independent determination of the astronomical unit, the mass of Venus, and the location of the primary ground tracking station.

Ground radio tracking stations are used to determine the spacecraft orbit. Owing to the high cost of such stations it is important to be able to visualize how the information from different stations combines to determine the orbit parameters. This paper will describe a three-dimensional model used successfully at the Jet Propulsion Laboratory to demonstrate how the tracking "geometry" of the different stations contributes to determination of target errors.

SPECIFIC EXAMPLES

Mars Orbiter Mission

The selection of the total flight path from parking orbit to the end of the mission is a complex job. We must understand the launch vehicle, the spacecraft subsystem requirements, and the scientific experiment requirements, as

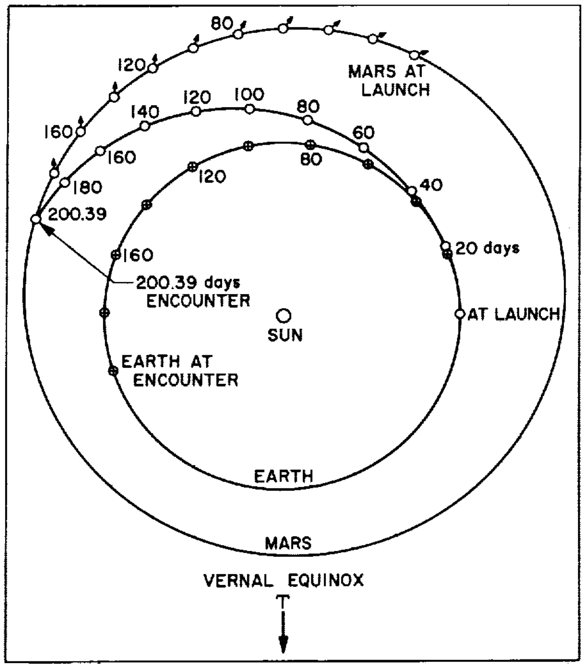


Figure 21-1.—A typical Earth-to-Mars transfer orbit.

well as basic celestial mechanics. We must know how the chosen flight path influences experiment payoff, spacecraft control accuracies, propulsion requirements, orbit determination accuracies, and engineering factors such as temperature control and power system requirements.

As a simple case in point, consider the mission of sending a spacecraft to orbit about Mars. Suppose that we want to photograph substantially all of the planet's surface in the course of five or more months in Martian orbit.

Figure 21-1 shows a typical Earth-to-Mars transfer orbit viewed from above the ecliptic plane. The compromises between launch vehicle payload capability, communications distance at encounter, and total flight time fairly tightly restrict the choice of transfer orbit for each opportunity. With the transfer orbit choices restricted, the spacecraft's approach velocity with respect to Mars is rather narrowly confined. This fact presents us with a potential problem with our orbiter.

Let us assume our initial choice is to investigate a polar orbiter, since this allows viewing the entire surface over a period of time. Figure 21-2, a simplified view of the polar orbiter as seen from above the ecliptic plane, has neglected the inclination of the Martian polar axis to the ecliptic plane in order to facilitate the discussion. The initial angle between the approach direction and the Sun-line will depend on the transfer trajectory. For practical orbiters the initial orbit plane direction will coincide with the approach direction. For a perfect polar orbiter, the angle Ω between the orbit plane and an inertial reference in the ecliptic plane is constant. A constraint to the effect that the orbit planè shall lie between $\pm\ \alpha$ degrees of the Mars Sun-line continuously over the first five months in orbit is imposed by the photographic lighting requirements and the spacecraft design under consideration. If the spacecraft approaches from the lower right in Figure 21-2, its initial orbit plane is in the acceptable sector. While the orbit plane remains fixed, the Mars-Sun direction Θ_s increases at about 0.5 deg/day due to Mars' travel in its orbit about the Sun, so that the shaded "acceptable sector" rotates to the left (counterclockwise) of the inertial reference line. For $\alpha=50°$ and for the most favorable approach angle, good lighting conditions are attainable over 100 deg/0.5 deg/day, or the 200 days required for the "acceptable sector" to rotate past the assumed fixed direction of the orbit

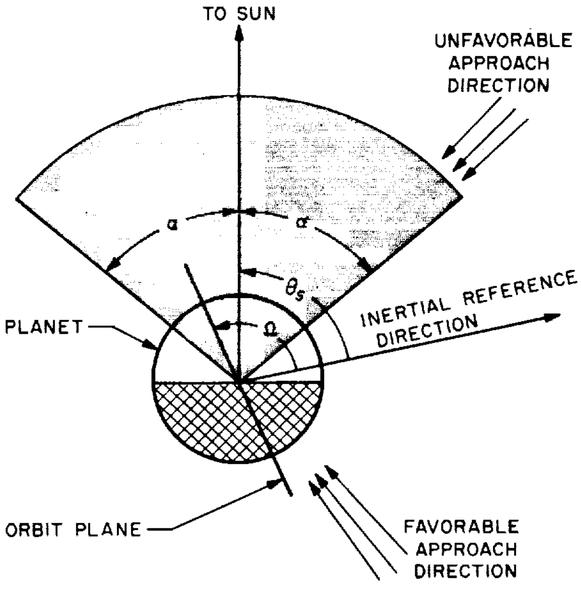


Figure 21-2.—Simplified view of orbit plane/Sun-line geometry. Acceptable lighting conditions occur when spacecraft orbit plane lies in the shaded sector of angular width 2α.

plane. For a very unfavorable direction such as indicated in the upper right of Figure 21–2 the lighting angle becomes unacceptable after only a few days.

If it is mandatory that transfer trajectories with a "favorable" approach direction be used, significant engineering penalties would result. Such penalties would include one to three months increase in flight time, greater communications distance at encounter, and possible reduction in allowable payload weight. Fortunately, we have overlooked an important factor. By using observations of the Martian satellites Phobos and Deimos, H. Struve, in 1911, accurately determined the "bulge" term in the gravitational field of Mars. The rate of rotation of the orbital plane about the polar axis is proportional to the "bulge" coefficient times the cosine of the orbit inclination to the Martian equatorial plane. By a moderate departure from a polar orbit and by guiding carefully in the terminal phase prior to establishing the spacecraft's orbit about Mars, we can cause the orbit plane's turn rate due to the "bulge" term to nearly equal the Sun-line's rotation rate and thus maintain acceptable lighting over a long period of time.

This simplified example only hints at the analytical problems connected with such an orbiter. Careful consideration must be given to the selection of the point of injection into Martian orbit, the orbit shape, and the influence of guidance errors. Another formidable problem would be the after-the-fact mapping and interpretation of the pictures taken from an orbiting spacecraft. The author is indebted to C. E. Kohlhase of JPL for his recognition and analysis of this situation.

Two-Way Doppler From Mariner 2 and Its Use

An example of one of the measurement types made possible by the spacecraft's existence is two-way doppler (Figure 21–3). The signal received in the spacecraft is shifted in frequency by the well-known doppler effect. The spacecraft then retransmits the signal which it has received. The signal received at the ground receiver has been further doppler-shifted by the radial velocity of the receiver with respect to the spacecraft. The difference in frequency between the received frequency and the current transmitter frequency is called the doppler tone.

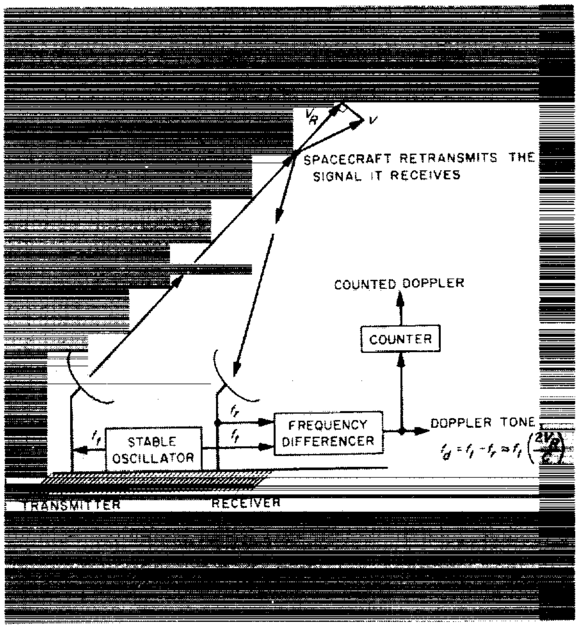


FIGURE 21–3.—Simplified two-way doppler configuration. The doppler tone f_d is a measure of the radial speed V_R.

In practice, the doppler tone is then averaged over an interval from 1 to 1000 seconds to obtain a data type known as counted two-way doppler. By combining such doppler measurements taken over an interval of time at several stations the spacecraft orbit may be reconstructed and its future course accurately predicted. On the Ranger series of lunar shots and on the Mariner 2 probe, currently on route to Venus, these measurements have been the primary source of information with which to determine the orbit.

Once the spacecraft orbit has been accurately estimated, the orbit can be altered by application of a small velocity increment using the midcourse correction rocket motor. Subsequent tracking of the spacecraft allows evaluation of the new orbit.

Figure 21–4 shows actual residuals obtained from tracking the Mariner 2 Venus probe from two of the Deep Space Instrumentation Facility (DSIF) stations operated by JPL for NASA. These residuals are the difference between the values of the observed two-way doppler and

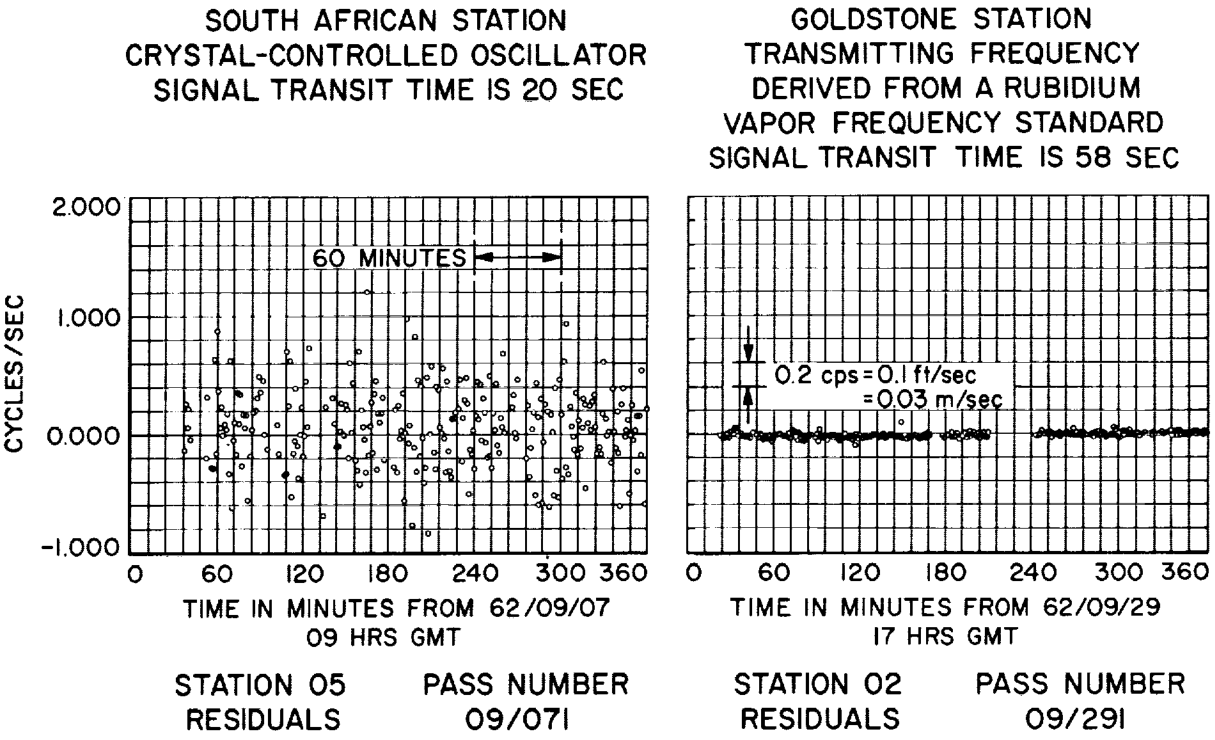


FIGURE 21-4.—Examples of residuals of counted two-way doppler taken on *Mariner 2* by 2 DSIF stations. Sample spacing is 60 sec; averaging time is 50 sec; 1 cps corresponds to 0.5 fps in radial velocity. Final corrections for station location errors have not been applied here.

those calculated on our current-best-estimate of the spacecraft orbit. The RMS error at the South African station is around 25 times that of the Goldstone station because the transmitter at Goldstone is more stable. Both accuracies are quite respectable; the RMS error at Goldstone is .01 ft/sec at a sampling rate of 60 per hour! Such high accuracy is hard to appreciate; by averaging 1 day's data we can estimate radial speed to the equivalent of .0004 ft/sec or 35 ft/day (in units of the speed of propagation). Our trajectory computation requires double-precision arithmetic and our orbit determination program must include factors normally considered negligible.

How does transmitter stability affect doppler accuracy? That we obtain an error if our transmitting frequency is not stable can be seen by considering $V_R=0$. The doppler tone should be zero, but will actually be the change in transmitter frequency during the time the signal takes to travel up to and back from the spacecraft. In the case shown here the signal transit time for South Africa is 20 sec; most of the noise in the residuals is due to a 0.5 parts in 10^9 shift in the ground transmitter frequency during that time. The Goldstone station shift in 58 sec seems surely below 1 part in 10^{11}, owing to the greater stability of its rubidium vapor frequency reference. Final evaluation of the stability of the Goldstone oscillator (Ref. 2) will not occur until Mariner nearly reaches Venus. At that point the signal round trip time is almost 7 minutes.

What will be the value of these precise measurements and how will they change celestial mechanics? First, they enable us to accurately predict the future course of the spacecraft, confirm the value of the astronomical unit, and perhaps measure the mass of Venus. A second and, seemingly, less likely benefit is the precise determination of the locations of the tracking stations on Earth.

One factor which complicates the astronomical unit determination based on Mariner 2 tracking data is the uncertainty of the spacecraft's effective reflecting area. The expected accelerations due to solar radiation pressure will cause about 3000 miles change in target error if completely neglected. In determining the AU we must separate the AU error from the area uncertainty.

The way the tracking station locations are determined from the doppler data may be understood by considering the spacecraft to be fixed with respect to the center of the Earth. The only doppler tone observed would be due to the rotating station's velocity component along the radial direction. The observed doppler tone at a station depends then on the latitude, longitude, and radius from the center of the Earth. Since we obtain measurements during many passes at a given station, and since the declination of the spacecraft changes during the mission, we may deduce the proper combination of station location errors required to minimize the errors in the residuals. Independent determination of the location of our primary tracking nation of the location of the two location coordinates perpendicular to the earth's spin axis is expected to an accuracy of about 20 meters by this method.

Visualizing Multistation Tracking Geometry

A typical lunar or interplanetary spacecraft is first placed into circular parking orbit by the booster vehicle. When the spacecraft-booster final stage combination has coasted to the right place, the final-stage rocket motor ignites and adds enough speed to intercept the target. The burnout of this rocket motor defines the point of injection into the transfer orbit to the target.

A network of tracking stations is required to track space probes after their injection into transfer orbit. Typically, these stations measure two angles defining the direction of the probe, range, range rate, or some combination of these. The purpose of these tracking stations is to assure that the future coordinates of the probe can be reliably estimated with sufficient accuracy to accomplish the mission.

Radio tracking stations are expensive and suitable sites are limited. To track all of our space probes from injection to injection plus a few hours would require a prohibitively large net of tracking ships and ground stations because of the wide variation of coasting time in parking orbit required to accomplish different planetary and lunar missions. A further complication is that the parking orbit inclination and coasting time are varied on a given launch day to compensate for variation of launch time within the allowed daily window. These variations cause the locus of possible injection locations to cover a large part of the Earth. The parametric study of the accuracy of determining the transfer orbit's elements on a spectrum of transfer orbits as a function of the number and location of the trackers, time from injection, measurement accuracies and types is a formidable task. It is not difficult to obtain the answer for any specific configuration, but it is difficult to generalize the results.

We know from our experience that, by combining the orbit parameter estimates independently obtained by several trackers, we can dramatically improve our knowledge of the orbit parameters, provided the combined "tracking geometry" is favorable. We have made some recent progress in developing methods of visualizing each tracker's "geometry." The insights gained will be useful in determining tracking station sites and accuracy specifications in an economical fashion.

In order to show how favorable "geometry" may be identified, consider the two-dimensional example shown in Table 21–II. In this example, the two coordinates of an archer's target error are independently estimated by observer A and observer B. A's location is such that he is able to estimate the X_2-coordinate well but is ten times as uncertain of the X_1-coordinate. This accuracy statement is contained in the covariance matrix Λ_A. The symbol E means the ensemble average. In the case at hand, observer B can estimate X_1 quite well but is less sure of X_2. Again this statement is contained in Λ_B. The minimum covariance estimate using both observer's estimates is obtained by the familiar expression on line 3 of Table 21–II. A convenient method for visualizing a 2 x 2 covariance matrix is to draw its "$1-\sigma$ error ellipse." The "$1-\sigma$ error ellipse" for observer A is given by the quadratic form $\bar{X}^T\Lambda_A{}^{-1}\bar{X}=1$. This ellipse will enclose 40 percent of the random occurences of $\bar{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ if $\bar{X}$ has a two-dimensional Gaussian probability distribution with covariance Λ_A. Such ellipses, or more generally ellipsoids in higher dimensions, are sometimes called *concentration* ellipsoids.

A particular property of the error ellipsoid of the combined estimator is useful here. The combined estimate's error ellipsoid is always interior (or tangent) to the error ellipsoid of each estimate taken separately. In the example here, each observer was uncertain by more than 10 units about the marksman's error, but the combined estimate's uncertainty is only one unit (line 4 of Table II). Here, it was evident that the "crossing" of these two narrow ellipses would closely determine the actual error.

Now that we have a satisfactory way of predicting when 2-dimensional estimates will combine favorably, let us see what happens when we push our luck to a 6-dimensional case. Can the geometry of each of the several stations tracking a lunar probe be usefully described in terms of each station's ability independently to predict the two components of target error?

Figure 21–5 illustrates that the answer is often "no." The solid ellipse on the right is obtained by the formally correct procedure described earlier, and the dotted ellipse is the weak bound obtained using the 2-dimensional approach. Figure 21–6 is a further example. Information from Station 3 improves the orbit of Station 5 more than the information of Station 4 does, even though the error ellipse of Station 4 is interior to that of 5. The reason for these failures to predict favorable "geometry" is that significant information has been suppressed.

However, we have found ways of successfully describing tracking station "geometry" in terms of 3-dimensional ellipsoids. The target error ellipses obtained from considering each station's ability to determine a properly chosen triplet of orbit parameters rather faithfully reproduces the final target error ellipses computed by the exact method. Different parameters are found to be appropriate to the lunar and interplanetary cases. The idea is most easily described for the interplanetary case (Ref. 1). Reference 3 describes coordinates which appear satisfactory for the lunar case. For interplanetary trajectories we use the three com-

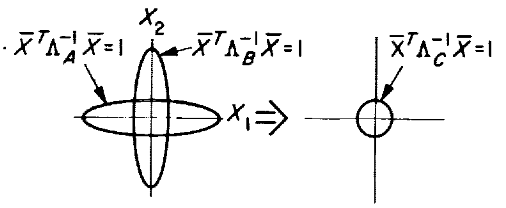
SITUATION	MATHEMATICAL EXPRESSION
1. OBSERVERS A AND B INDEPENDENTLY ESTIMATE THE 2 COMPONENTS OF AN ARCHER'S MISS	$\bar{X}_A = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}_A \quad , \quad \bar{X}_B = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}_B$
2. A'S UNCERTAINTY IS 10 UNITS IN X_1 AND 1 UNIT IN X_2 WITH NO CORRELATION. B'S UNCERTAINTY IS 10 UNITS IN X_2 AND 1 UNIT IN X_1 WITH NO CORRELATION	COVARIANCE $\bar{X}_A = \begin{pmatrix} (10)^2 & 0 \\ 0 & (1)^2 \end{pmatrix} = \Lambda_A = E\left(\bar{X}_A \bar{X}_A^T\right)$ COVARIANCE $\bar{X}_B = \begin{pmatrix} (1)^2 & 0 \\ 0 & (10)^2 \end{pmatrix} = \Lambda_B$
3. THE "BEST" COMBINATION OF THE TWO ESTIMATES IS OBTAINED BY RELYING ON EACH SOURCE FOR THE PART IT ESTIMATES BEST	$\bar{X}_C = \left(\Lambda_A^{-1} + \Lambda_B^{-1}\right)^{-1} \left(\Lambda_A^{-1} \bar{X}_A + \Lambda_B^{-1} \bar{X}_B\right)$
4. THE COMBINED ESTIMATE IS BETTER THAN EITHER ESTIMATE ALONE	$\Lambda_C = \left(\Lambda_A^{-1} + \Lambda_B^{-1}\right)^{-1} = \begin{pmatrix} 1.01 & 0 \\ 0 & 1.01 \end{pmatrix}^{-1} = \begin{pmatrix} 0.99 & 0 \\ 0 & 0.99 \end{pmatrix}$
5. THE "ERROR ELLIPSE" OF A'S ESTIMATE IS SMALL IN THE X_2 DIRECTION WHILE THE "ERROR ELLIPSE" OF B'S ESTIMATE IS SMALL IN THE X_1 DIRECTION. THE "GEOMETRY" IS FAVORABLE	 $\bar{X}^T \Lambda_A^{-1} \bar{X} = 1$ X_2 $\bar{X}^T \Lambda_B^{-1} \bar{X} = 1$ X_1 $\Rightarrow$ $\bar{X}^T \Lambda_C^{-1} \bar{X} = 1$

TABLE 21–II.—Visualizing How Favorable "Geometry" Can Give Dramatic Improvement When Combining Two Estimates.

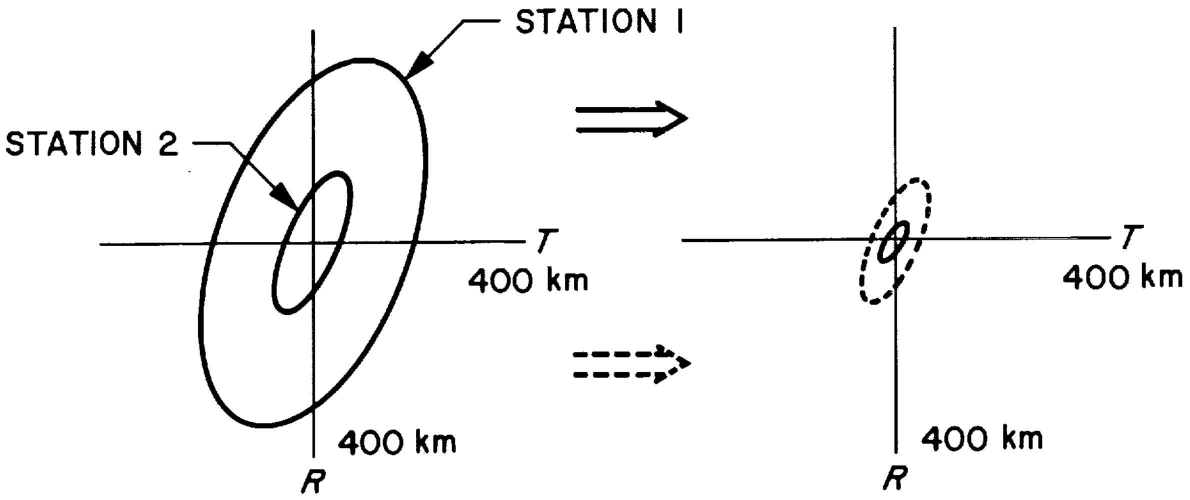


FIGURE 21-5.—Comparison of actual target error ellipse (solid) with weak bound (dashed), obtained by considering only each station's 2-dimensional error ellipse.

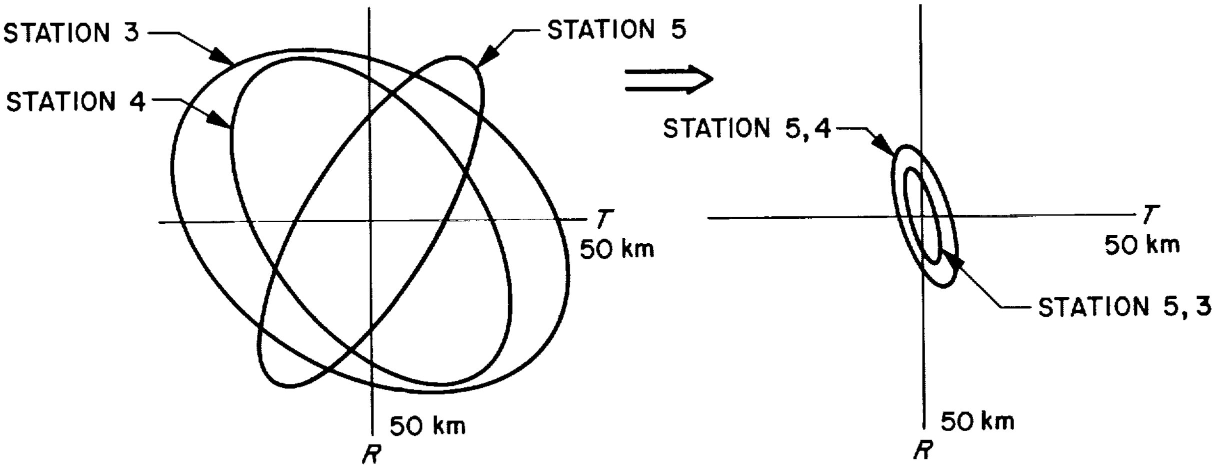


FIGURE 21-6.—Comparison showing that the 2-dimensional ellipse of target errors for each station alone fails to suggest the favorable "geometry" actually existing.

ponents of the geocentric hyperbolic excess velocity vector as parameters. Any other three parameters specifying the position at a given time may be considered to complete the set.

Instead of associating a 6 x 6 covariance matrix with each station's tracking geometry, we ignore all but the 3 x 3 covariance matrix of errors in the hyperbolic excess velocity. The "geometry" of each station is considered to be described by the three-dimensional ellipsoid in these parameters. The terms which are ignored do not significantly alter the target error ellipsoid. This technique works for interplanetary trajectories because the target error effect of velocity errors at entry into the Sun-centered phase of the flight dominates the position errors due to the great time available for them to propagate.

CONCLUSION

The three subjects chosen for discussion are a fair sampling of current work in spacecraft flight studies. The great variety of possible flight paths and the interaction between flight path and mission objectives require that the space systems analyst be well versed in both celestial mechanics and engineering. Precise descriptions of spacecraft motion in various situations will have to be understood by many people in order to plan and carry out future missions. New tools of analysis are required to deal with both the new measurements and new applications.

The controllable spacecraft is a powerful new instrument with which we shall improve our description of the world in which we live. If this paper has aroused your interest and given you the feeling that the surface of the subject has only been scratched, it will have met its objective.

REFERENCES

1. CLARKE, V. C. JR., et al, *Earth-Venus Trajectories, 1964*, TM 33–99, Vol. 1, Part A.
2. *Evaluation of Atomic Oscillator Performancc*, Space Programs Summary No. 37–17, Vol. III, Jet Propulsion Laboratory, P. 30.
3. THORNTON, T. H., JR., *A Study of Error Sensitivity for Lunar Trajectories*, Space Programs Summary No. 32–15, Vol. IV, Jet Propulsion Laboratory, pp. 4–7.

22. Space Flight Optimization

By William G. Melbourne

Dr. WILLIAM G. MELBOURNE, *Research Group Supervisor in Systems Analysis Research for Jet Propulsion Laboratory, was born and educated in Los Angeles, California. The University of California at Los Angeles conferred the degree of A.B. in Astronomy-Physics in 1954; California Institute of Technology conferred the degree of Ph. D. in Astrophysics in 1959. Dr. Melbourne is a member of ARS and Sigma Xi.*

INTRODUCTION

In recent years a substantial effort has been expended on the development and adaptation of systems optimization techniques to space flight problems. With the advent of complex space vehicles, optimization theory has been applied to the problems of vehicle design, to the control policies under which such systems operate, and to the design of missions for which these systems are intended. Thus, in vehicle design analysis one strives for an optimal configuration subject to the many and diverse engineering constraints which are imposed. In this case, the word "optimum" bears a rather complex connotation since the ultimate configuration will evolve from a consideration of the interplay between such factors as maximum payload, reliability, redundancy, stability, state of the art, etc. An optimal control policy or an optimal mission design, on the other hand, generally accomplishes a more definitive objective such as a minimum propellant expenditure, but as before, it is subject to the fulfillment of certain constraints. The thrust program of a space vehicle, for example, may be optimized on the condition that the thrust magnitude lie within certain bounds or that the thrust steering program be limited so that the structural limits of the vehicle are not exceeded. Because the complexity of these problems generally defies an intuitive attack and because the system performance is in many cases very sensitive to perturbations in design and policy, the use of optimization procedures in one form or another has become a mandatory part of systems analysis.

The scope of this paper is mainly restricted to flight analysis, with particular emphasis on vehicle performance. Consequently, it will be assumed that a well-defined goal for the optimization process exists. A measure of the degree to which this goal is achieved is usually provided by a criterion of optimization. Frequently, this criterion is that some function, often called the payoff function, of the state variables and parameters of the problem should be an extremum—that is, either a maximum or a minimum. Such quantities as maximum payload, minimum fuel expenditure, maximum satellite altitude, maximum range, minimum time, minimum target miss, etc., might each be a goal of the optimization process. The payload of a vehicle, for example, may be considered as the difference between the vehicle weight after accomplishing the mission and the residual components of the vehicle such as supporting structures, communications, power and altitude control equipment, etc. Maximizing the payload, therefore, will involve an optimization of both the thrust program and the mission design for minimum propellant expenditures as well as an optimal design of the components of the vehicle.

In order to formalize the discussion, let us provide definitions for three types of quantities appearing in systems optimization problems. The quantities appearing in these problems may be categorized as state variables, control variables, and system parameters. As examples of each of these types, consider a vehicle which is assumed to be a point mass traveling through space. The state variables of this system are the three position coordinates of the vehicle, the three velocity coordinates, and the instantaneous mass of the vehicle. The state variables are generated from a set of differential equations which, in this example, are simply Newton's equations of motion and a continuity equation relating the propellant flow to the mass loss rate of the vehicle. The control variables for this example might be the thrust magnitude of the vehicle and a set of angles defining the thrust direction. The system parameters are constants describing certain properties of the problem. Such parameters might be the exhaust velocity of the propulsion system or a prespecified time of thrust termination. For advanced systems such as ionic propulsion systems, these parameters might be the values of the exhaust velocity and the size of the powerplant carried by the vehicle.

GOALS OF THE OPTIMIZATION PROCESS

What are the broad objectives of a systems optimization process? The optimization process should provide control policies, parameter configurations, and mission designs which are optimal and from which the following kinds of information are available. First, the optimization process yields extremal values of the payoff function for various ranges of mission conditions. For examplé, one might obtain, for a given space vehicle, the minimal fuel expenditure for a particular interplanetary mission as a function of flight time. Second, the degradation in performance as measured by the departure of the payoff function from its extremal value which results from the use of nonoptimal control policies, parameter configurations or mission design may be obtained. Furthermore, the effects of imposing additional constraints on state variable, control variables and system parameters with a subsequent reoptimization consonant with these constraints may be assessed. An example of a nonoptimal interplanetary mission design is the launching of a space vehicle toward the planet on a date requiring more than the minimum propellant expenditure. A knowledge of the variation in propellant requirement with launch date is obviously a necessity in mission planning exercises. Thus, the rate of degradation in performance or sensitivity to nonoptimal operations or constraints is important. It has been found that the return leg of Mars round-trip trajectories frequently has a perihelion distance which is less than 1 astronomical unit. Such a trajectory may be objectionable because of the increased solar radiation density, and it may be necessary to impose a constraint on the distance of closest approach to the Sun. This kind of state variable constraint will increase the propellant requirements and it will be necessary to determine the penalty caused by the additional constraint and the resulting modifications to the original trajectory.

OPTIMIZATION TECHNIQUES

The resurgence of optimization theory in systems problems naturally has been accompanied by a vigorous development of analytical and numerical techniques for formulating and solving these types of problems. For extremal problems there are three principal techniques which have gained currency in recent times. The classical method is, of course, the calculus of variations in which the optimization is accomplished by satisfying a set of conditions appearing mainly as differential equations. The calculus of variations had its origin in the 17th century with the work of the Bernoulli brothers on the brachistochrone problem. From this point, a series of contributors highlighted by such names as Euler, Lagrange, Legendre, Jacobi, Weierstrass, Hilbert, Mayer, Bolza, and Bliss honed the calculus of variations into a moderately complete discipline as summarized in the works of Bolza (Ref. 1) and Bliss (Ref. 2). The researches of Valentine (Ref. 3) opened the way for the application of the calculus of variations to problems containing bounded control variables. Early in the last decade the work of several investigators, notably Cicala (Ref. 4) and Hestenes (Ref. 5) rendered the calculus of variations into a more tractable form for flight analysis problems. Recently, the work of Pontryagin (Ref. 6) in optimal control theory leading to the "Maximum Principle" has strengthened the calculus of variations so that it is applicable to a wider class of problems, such as, for example, systems possessing discretely varying control variables.

As we shall see, the calculus of variations formulation generally leads to a high-order system of nonlinear first-order differential equations. Except in the simplest of problems, numerical methods of integration must be used, although one or two constants of integration usually are available analytically to reduce the order of the system accordingly. When numerical methods are necessary, one is generally confronted with the mixed or "two-point" boundary-value problem owing to the fact that boundary conditions are specified at both the initial and final points of the solution. In order numerically to generate a solution one must have on hand initial values for all the variables being integrated, and so one guesses values for the unspecified initial conditions, integrates the equations, and checks the agreement of the specified final conditions with the corresponding integrated condition. This leads to a trial and error process which is time-consuming and which presents the principal difficulty in solving systems optimization problems with the calculus of variations.

There are two iterative methods for surmounting the two-point boundary-value problem. The direct difference method relating final values to initial values is often used in conjunction with an interpolation scheme such as the Newton-Raphson or Runge-Kutta methods to attempt to null the difference between the specified and integrated final values. Because of departures from linearity, this process must usually be repeated until satisfactory convergence is attained. The adjoint method (Ref. 7) provides the second approach, and it will be briefly described for a relatively simple form of boundary conditions. Suppose one has a system of differential equations given by

$$x_i = g_i(\vec{x}, t) \qquad i=1, \ldots, n \tag{1}$$

with boundary conditions such that the first r components of $\vec{x}$ are specified initially $(t=t_0)$ and the remaining n-r components are specified at the final point $(t=t_1)$. Then taking the first variation of Eq. (1) holding time fixed, one obtains

$$\frac{d}{dt}(\delta x_i) = \frac{\partial g_i}{\partial x_j}\,\delta x_j, \qquad i=1, \ldots, n \tag{2}$$

where the summation rule is employed. The adjoint variables are defined by

$$\dot{\lambda}_i = -\frac{\partial g_j}{\partial x_i}\,\lambda_j \qquad i=1, \ldots, n \tag{3}$$

and it follows from Eq. (2) and (3) that

$$\vec{\delta x} \cdot \vec{\lambda}\,\Big|_{t_0}^{t_1} = 0 \tag{4}$$

If one now defines

$$\lambda_i^{(k)}(t_1) = \delta_{ik} \qquad k=r+1, \ldots, n \tag{5}$$

there results

$$\delta x_k(t_1) = \vec{\lambda}^{(k)} \cdot \vec{\delta x}\,\Big|_{t_0} \qquad k=r+1, \ldots, n \tag{6}$$

Using an initial solution of Eq. (1) one integrates Eq. (3) backwards n-r times with k successively taking on values from $r+1$ to n. Then, using the specified values minus the values from the initial solution for $\delta x_k(t_1)$ and the matrix coefficients $\lambda_i^{(k)}(t_0)$ generated by Eq. (3), one may invert Eq. (6) and solve for the $\delta x_k(t_0)$, $(k=r+1, \ldots, n)$, which form a set of corrections to the unspecified initial conditions. The success of both of these methods will depend, of course, on the degree of linearity which holds between the initial and final values of the variables. Experience has shown (Ref. 8) that at least in the small, these methods generally succeed even with systems of equations of order 12 and higher. Both methods have their advantages and are about equal in computational time. The adjoint method does not suffer a loss of significant figures which sometimes occurs in the direct difference method; on the other hand, the difference method is usually more versatile and does not require the solution of the adjoint equations.

The direct method of gradients or steepest descents is the second optimization technique which is in active use today. The application of this method to variational problems was first developed by Hadamard and later by Courant (Ref. 9, 10). More recently, the method has been augmented and applied to flight mechanics problems by Kelley (Ref. 11), Bryson (Ref. 12), and others. This method obviates the two-point boundary-value problem by employing an iterative process in which each successive solution is forced to satisfy the boundary conditions. One commences with an initial solution which satisfies the boundary conditions and which has certain adjustable parameters depicting the functional forms of the control variables of the problem. These parameters are then adjusted, consonant with the boundary conditions, in directions which effect the greatest change in the payoff function, that is, along the directions of steepest descent. The criteria for adjusting the control variable parameters are updated from the subsequent solution, and the whole process is repeated until convergence is attained. One of the chief attractions of the gradient method is that it generally converges to the optimal solution even though the initial trial solution is significantly nonoptimal. In view of this, there have recently been developed hybrid methods in which the gradient method is used to obtain a nearly converged solution, after which the calculus of variations is used to complete the iteration process.

The third optimization technique is drawn from the field of dynamic programming in which the principle of optimality (Ref. 13–15) is applied to the payoff function in order to obtain a functional recurrence relation suitable for computational purposes. In this method the continuous process is reduced to a set of discrete processes or stages. The principle of optimality states that "an optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision." At each stage a search is conducted over the discrete control vector space of that stage in order to isolate the optimal policy for a given state subject to the fulfillment of the optimality principle for the subsequent stages. By this optimality principle, the search for an optimal policy, as opposed to a direct enumeration process, is made tractable. Ironically, the existence of constraints on either the state or control variables facilitates the search process, and the two-point boundary value problem does not exist since boundary conditions are simply treated as constraints on the initial and final stages of the process. The solving of optimum trajectory problems by dynamic programming has been hampered, nevertheless, by the large dimension of the grid of stored quantities which is necessary for the search process. An n-dimensional trajectory problem requires at least a 2n-dimensional grid of stored quantities. Thus, for a two-dimensional trajectory problem, the number of stored quantities is of the order of N^4 where N is the number of stored quantities along one axis of the grid. Generally, N will be greater than 10 for reasonable accuracy, in which case, both from the standpoint of storage and computational time, the problem is not practicable for modern high-speed computers. Some progress has been made in reducing the storage problem, notably by the use of polynomials to approximate the stored data, and successive approximation.

Calculus of Variations in Space Flight Analysis

In this section we present a formulation of the calculus of variations which is readily adaptable to flight mechanics problems. For simplicity in the algebra it will be assumed that all the initial conditions are specified and that the quantity being minimized (or maximized) is a function of the state variables evaluated at the final point and the system parameters. In this case, the payoff function J becomes

$$J=J[y_1(t_1),\ldots,y_n(t_1),\kappa_i,\ldots,\kappa_r,t_1] \qquad (7)$$

and the problem is to determine the conditions on the state variables $y_j(t)$, the control variables $u_i(t)$ $(i=1,\ldots,m)$, and the system parameters κ_i, which provide an extremal in J. The state variables are subject to the differential equation constraints given by

$$G_i=\dot{y}_i-f_i(\vec{y},\vec{u},\vec{\kappa},\vec{t})=0 \qquad i=1,\ldots,n \quad (8)$$

and the control variables and system parameters are constrained by algebraic equations of the form

$$G_{n+i}(\vec{y}, \vec{u}, \vec{\kappa}, \vec{t})=0 \qquad i=1, \ldots, p \tag{9}$$

In addition, the state variables are subject to a set of boundary conditions at the final time

$$A_i(\vec{y}, t_1)=0 \qquad i=1, \ldots, q\leq M \tag{10}$$

In order to handle inequality constraints (Ref. 3) on certain control variables of the form

$$u_{i\,\min}\leq u_i\leq u_{i\max} \tag{11}$$

one may increase the dimensionality m, of the control vector space by defining the quantities $u_{m+i}(t)$ to be real variables given by

$$G_{n+i}=u^2_{m+i}-(u_{i\,\max}-u_i)(u_i-u_{i\,\min})=0 \tag{12}$$

which will guarantee the satisfaction of Eq. (11). A similar formalism can be applied to the system parameters. For dealing with inequality constraints on state variables the reader is referred to a recent paper by Dreyfus (Ref. 16). It should be pointed out that the system parameters may also be treated as state variables (Ref. 4) generated by the equations

$$\dot{\kappa}_i=0 \tag{13}$$

but, for convenience, we will adhere to the formulation as presented.

The foregoing formulation is essentially the Mayer problem (Ref. 1, 2) of the calculus of variations, and upon applying the theory one obtains as necessary conditions for an extremal value of J the Euler-Lagrange equations

$$\frac{d}{dt}\left(\frac{\partial F}{\partial \dot{y}_i}\right)-\frac{\partial F}{\partial y_i}=0 \qquad i=1, \ldots, n \tag{14}$$

$$\frac{\partial F}{\partial u_i}=0 \qquad i=1, \ldots \tag{15}$$

and

$$\int_{t_0}^{t_1} \frac{\partial F}{\partial \kappa_i}\, dt=0 \qquad i=1, \ldots, r \tag{16}$$

where

$$F=\sum_{j=1}^{n+p} G_j\lambda_j(t) \tag{17}$$

and the $\lambda_j(t)$ are Lagrange multipliers. A further necessary condition for a local minimum in J which is useful in flight analysis is the Weierstrass E-function (Ref. 1, 2), which for this formulation is

$$E=\sum_{j=1}^{n} (\dot{y}_j-\dot{y}^*_j)\,\frac{\partial F}{\partial \dot{y}_j}\geq 0 \tag{18}$$

where the $\dot{y}^*_j$ are any permissible departures from the optimal values $\dot{y}_j$ due to nonoptimal but permissible values of the control variables. Using Eq. (8) and (17), it is easily seen that the Weierstrass E-function is equivalent to the condition

$$H=\underset{\vec{u}\in\vec{U}}{\mathrm{Max}}\left\{ {}_{j=1}^{n} \quad \lambda_j f_j(\vec{y}, \vec{u}, \vec{\kappa}, t)\right\} \tag{19}$$

which is Pontryagin's maximum principle (Ref. 6). The quantity $\vec{U}$ is the space of permissible values for $\vec{u}$ as imposed by Eq. (9). This equivalence is noted here because Pontryagin's work is applicable to a wider class of control variables such as discrete control variables, while Eq. (18) was derived by Weierstrass under more stringent continuity assumptions. In this manner, as pointed out earlier, the classical methods may be strengthened by the inclusion of the maximum principle.

Finally, there are the boundary conditions to be considered. Eq. (14–18) serve, essentially, to determine the optimal values of the control variables and the system parameters. From the simultaneous solution of the relations

$$dJ=\nabla J\cdot \vec{d}y(t_1)+\frac{\partial J}{\partial t}\, dt_1=0 \tag{20}$$

$$dA_i=\nabla A_i\cdot \vec{d}y(t_1)+\frac{\partial A_i}{\partial t_1}\, dt_1=0 \qquad i=1, \ldots, q \tag{21}$$

and the general transversality condition obtained from the calculus of variations

$$[\nabla\dot{y}F\cdot \vec{d}y-Hdt]|_{t1}=0 \tag{22}$$

one is provided with the requisite number of boundary conditions to obtain the complete optimal solution. One of the important results from Eq. (20–22) is that if certain terminal quantities are undetermined by Eq. (10), there results a corresponding transversality condition, in effect, for each undetermined quantity. Satisfying this transversality expression yields an extremal in J with respect to the corresponding undetermined quantity. It is also easy to obtain from Eq. (20–22) the first variations in J with respect to the final values of the state variables or the system parameters.

A FLIGHT OPTIMIZATION PROBLEM

The applications of the calculus of variations to flight mechanics problems has been extensive in the last decade, and the general theory of optimal flight paths has been developed for ballistic vehicles (Ref. 17–20) and for advanced propulsion systems such as power-limited vehicles (Ref. 21–23). For illustrative purposes, we now discuss the problem of optimizing the trajectory of a power-limited system under various constraints. The vehicle is assumed to be a point mass travelling in a vacuum and subjected to a conservative force field. The constraining equations for such a system are given by Newton's equations

$$\dot{\vec{v}}+\nabla V-\frac{\beta}{c\mu}\overrightarrow{\alpha_p l}=\vec{0} \tag{23}$$

$$\vec{r}-\vec{v}=\vec{0} \tag{24}$$

and the power-limited constraint relating vehicle mass loss rate $\dot{\mu}$ to propulsion parameters

$$\dot{\mu}+\frac{\beta}{c^2}\alpha_p=0 \tag{25}$$

The state variables are position and velocity coordinates r and v, and the normalized vehicle mass $\mu[\mu(t_0)=1]$. The control variables are the direction cosines of the thrust vector $\vec{l}$ and α_p is a normalized power parameter having the value 1 during propulsion periods and 0 during coasting periods. The control variable constraints may be written as

$$|\vec{l}|-1=0 \tag{26}$$

$$\alpha_p=0,1 \tag{27}$$

The system parameter in this problem is β, which is twice the kinetic power in the rocket exhaust divided by the initial vehicle mass. If the exhaust velocity c is held fixed it is also a system parameter; however, it may also be a control variable through which the magnitude of the thrust is varied. The thrust acceleration a is given by

$$a=\frac{\beta}{c\mu}\alpha_p \tag{28}$$

and if this is combined with Eq. (25) and integrated, one obtains the rocket equation for power-limited flight.

$$\frac{1}{\mu_1}=1+\frac{1}{\beta}\int_{t_0}^{t_1}a^2dt \tag{29}$$

Let us now find the optimal policies for which the final mass μ_1 is maximized. Since this is equivalent to minimizing $\int_{t_0}^{t_1}a^2dt$ and since this quantity is essentially independent of the propulsion system parameters (Ref. 23), the payoff function will be taken as

$$J=\int_{t_0}^{t_1}a^2dt=\beta\left(\frac{1}{\mu_1}-1\right) \tag{30}$$

Upon applying the calculus of variations to this problem, one obtains the following optimality conditions on the control variables: The optimal direction of thrust is given by

$$\vec{l}=\vec{\lambda}/\lambda \tag{31}$$

where $\vec{\lambda}$ is the vector sum of the three orthogonal Lagrange multipliers associated with Eq. (23) and is generated by the 6th-order system of differential equations

$$\ddot{\vec{\lambda}}+(\vec{\lambda}\cdot\nabla)\nabla V=\vec{0} \tag{32}$$

In addition, upon defining the switching function L to be generated by

$$\dot{L}=\vec{l}\cdot\dot{\vec{\lambda}}/\mu \tag{33}$$

The optimal conditions for coast and propulsion are given by

$$\left\{\begin{matrix} L>0, & \alpha_p=1 \\ L<0, & \alpha_p=0 \end{matrix}\right\} \tag{34}$$

If c is considered as an unbounded control variable then one finds that it is given by

$$c=k/\mu\bar{l}\cdot\bar{\lambda} \tag{35}$$

and the thrust acceleration by

$$a=\beta\bar{l}\cdot\bar{\lambda}/k \tag{36}$$

where the constant k is determined by boundary conditions. If c is a constant, it may be shown that its optimal value as a system parameter occurs when the condition

$$\int_{t_0}^{t_1}\alpha_p\left(2L-\frac{\bar{l}\cdot\bar{\lambda}}{\mu}\right)dt=0 \tag{37}$$

is satisfied. For those cases in which the force field potential V is explicitly independent of time, it may be shown that a constant of integration results and is given by

$$H=\frac{\beta L}{c}\alpha_p-\dot{\bar{\lambda}}\cdot\dot{\bar{r}}-\vec{\lambda}\cdot\nabla V \tag{38}$$

where H is the Hamiltonian constant appearing in Eq. (19). For further details the reader is referred to Ref. 22 and 23. The system of differential equations to be numerically integrated is of the 15th order if Eq. (37) is included· A slight reduction in order can be obtained from additional constants of integration when they are available (Ref. 22, 23). However, Eq. (38) is usually not suitable for numerical integration because of the $\dot{\bar{\lambda}}\cdot\dot{\bar{r}}$, but it does serve as a check on the accuracy of the integration.

This formulation has been applied to an interplanetary rendezvous mission from Earth to Mars. Therefore, both the initial and final positions and velocities are matched with the heliocentric positions and velocities of the planets. An inverse square force field model using the mass of the Sun was used and the planets were assumed massless. In these trajectories neither the position on the orbit of Mars (true anomaly) nor the transfer angle from the Earth to Mars were specified; consequently, two transversality conditions arise, and were satisfied instead (Ref. 22). Accordingly, the values of J which result are associated with trajectories corresponding to launch and arrival dates for which the Earth-Mars planetary configuration is optimum. Figure 22–1 exhibits the variation of J with heliocentric flight time for three types of thrust programs and for two sets of boundary conditions. Since the transversality conditions guarantee only local extremal values in J, it is also possible to generate local maxima as well. Although the curves in Figure 22–1 correspond to trajectories with optimum launch and arrival dates, the upper set corresponds to synodic years in which the position of Mars on its orbit is least optimum (e.g., 1964) while the lower set corresponds to the optimum orbital position (e.g., 1971). Thus, these two sets of curves bound the values of J which are available in any synodic year, provided that optimum launch dares are used within that year.

The three curves within each set reflect the use of thrust programs with different constraints. The best performance is obtained from the variable thrust program in which c is an unbounded control variable and is depicted by the No. 1 curves. In the No. 2 curves, c was fixed, but its value and the resulting length of coast were chosen so that Eq. (37) was satisfied yielding a minimum in J with respect to c. Finally, in the No. 3 curves, no coast was permitted and the fixed value of c was determined by the boundary conditions. These are also "minimum time" trajectories for a given initial acceleration. From a study of results such as Figure 22–1, one may accurately assess the degradation in performance which results from control variable constraints and departures from optimal mission design. Figure 22–2 exhibits a 160-day Earth-Mars rendezvous trajectory generated by using an optimal constant thrust program with optimum coast. The arrows indicate the direction of thrust.

We now turn to a more restricted thrust program in which the direction of thrust is constrained to prespecified discrete values $\vec{l}_i$, that is,

$$\vec{l}=\vec{l}_i \qquad i=1,\ldots,r \tag{39}$$

The optimal control for this thrust program will be established and the criteria for optimizing these prespecified thrust directions will be developed. This constant-attitude thrust program has a practical importance since it is probably the simplest program which can be executed by a Sun-oriented space vehicle.

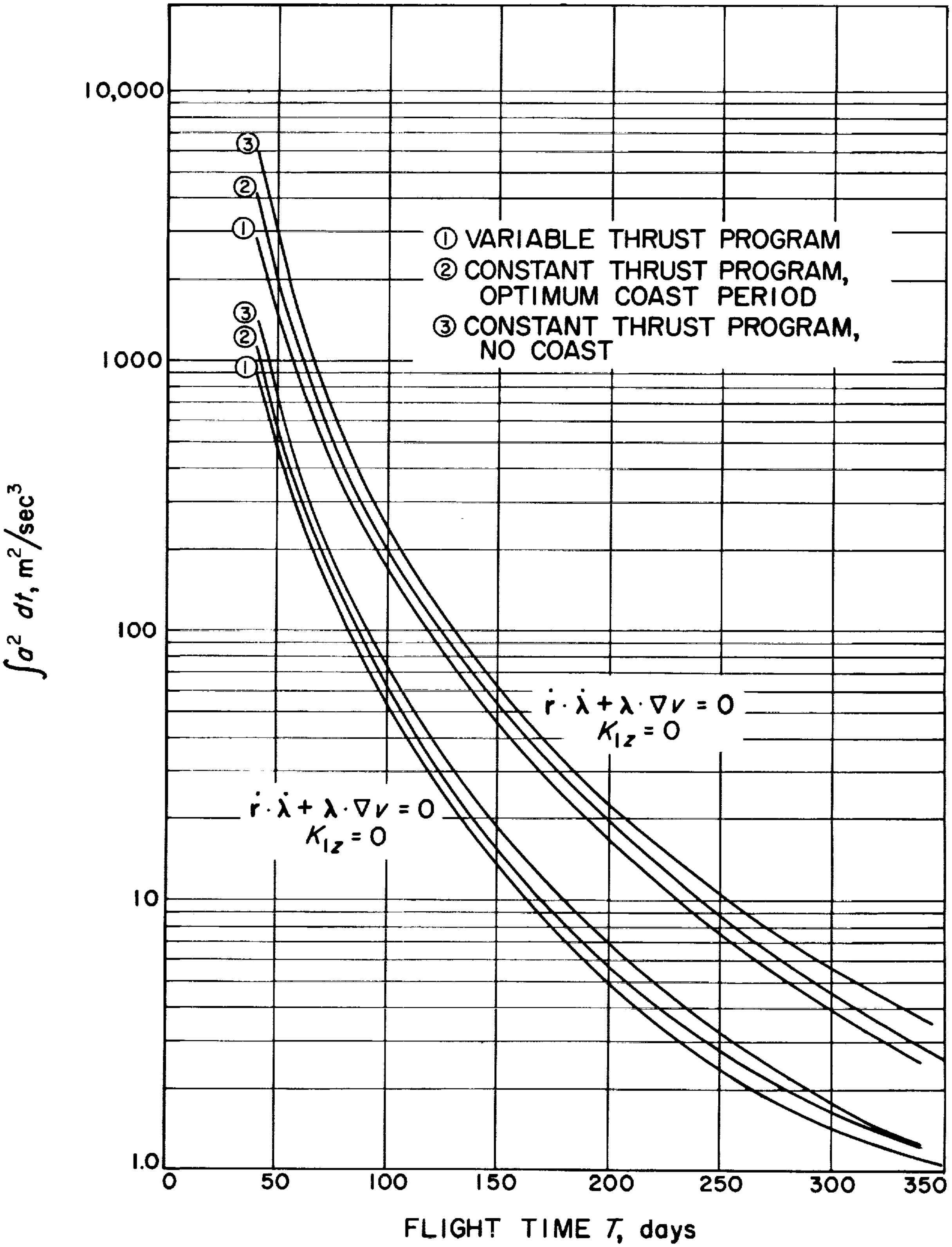


FIGURE 22-1.—J versus flight time for Earth-Mars rendezvous missions.

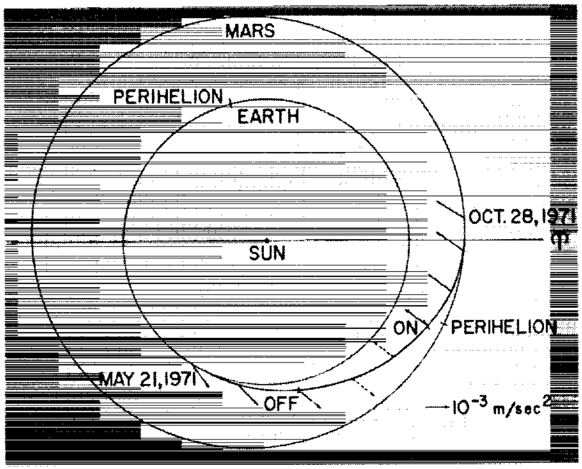


FIGURE 22–2.—160-day optimum rendezvous trajectory with constant thrust program, optimum variable direction.

Upon applying the optimization theory to this problem, one finds that Eq. (32–38) still holds. By use of the maximum principle, however, the conditions for the optimal choice of $\vec{l}$ is given by

$$(\vec{l}-\vec{l}^*)\cdot\vec{\lambda}\geq 0 \qquad (40)$$

Thus, the optimal thrust direction at any point along the trajectory is that direction taken from the discrete set $\vec{l}_i$ which is most nearly parallel to $\vec{\lambda}$. This result is quite expected since, in the unconstrained program, the optimum thrust direction is, by Eq. (31), along $\vec{\lambda}$.

There remains the problem of optimizing the values of the prespecified thrust direction $\vec{l}_i$ which, in this case, is a system parameter optimization process. It may be shown from Eq. (16) that the condition for optimum $\vec{l}_i$ is given by

$$\int_{t_0}^{t_1} a(\vec{\lambda}x\vec{l}_i)\alpha_{l_i}dt=0 \qquad i=1, \ldots, r \qquad (41)$$

where the quantity α_{l_i} has the value 1 during those phases where $\vec{l}$ has the value $\vec{l}_i$ and zero otherwise. This condition is also quite expected, since in the variable direction program, where $\vec{l}_i$ varies continuously, the integrand of Eq. (41) is zero at every point along the trajectory.

A series of two-dimensional Earth-Mars rendezvous trajectories of the type corresponding to the lower No. 2 curve of Figure 22–1 has been generated using this constant-attitude thrust program. In this case, two prespecified thrust directions relative to the heliocentric radius vector were allowed. These directions were denoted by the angles Γ_1 and Γ_2, as indicated in Figure 22–3. The choice of thrust direction at any point on the trajectory is determined by Eq. (40), and two directions themselves have been optimized by satisfying Eq. (41); thus J possesses a local minimum with respect to Γ_1 and Γ_2. Figure 22–4 shows a 160-day trajectory using this program, and the similarity with Figure 22–2 should be noted. Figure 22–5 shows the variation of the optimal values of Γ_1 and Γ_2 with flight time. Figure 22–6 shows the percentage excess in J which results from the use of the constant-attitude program instead of the optimally directed program.

Finally, it is interesting to investigate the sensitivity of J to departures of the Γ_i from their optimal values. It may be shown that

$$\frac{\partial J}{\partial \Gamma_i}=\frac{a_0}{\mu_1(\bar{l}\cdot\bar{\lambda}-\mu L)_{t_1}}\int_{t_0}^{t_1} a(\vec{\lambda}x\vec{l})\alpha_{\Gamma_i}dt \qquad (42)$$

Figure 22–7 exhibits both the variation in J and $\partial J/\partial\Gamma_1$ with Γ_1 for a 160-day trajectory. This figure suggests that the sensitivity of J to the choice of Γ_1 in the vicinity of the optimal value is not particularly critical. Similar considerations also apply to Γ_2.

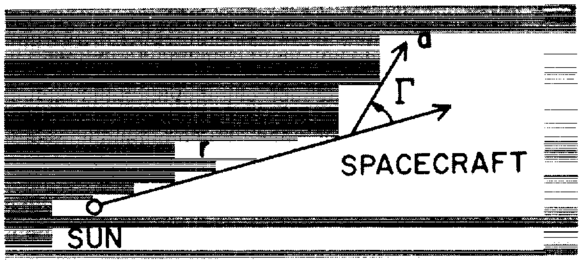


FIGURE 22–3.—The constant attitude thrust direction.

In summary, it has been shown that the constant-attitude thrust program with optimized thrust directions is competitive in vehicle performance with the optimal variable direction program. By the use of two optimized thrust

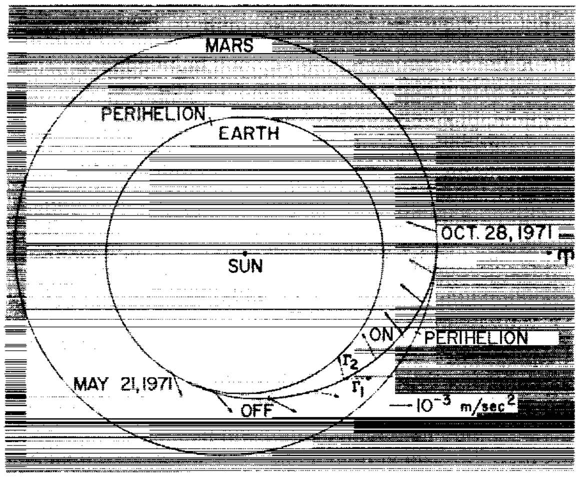


FIGURE 22-4.—160-day optimum rendezvous trajectory with constant thrust program, constant attitude.

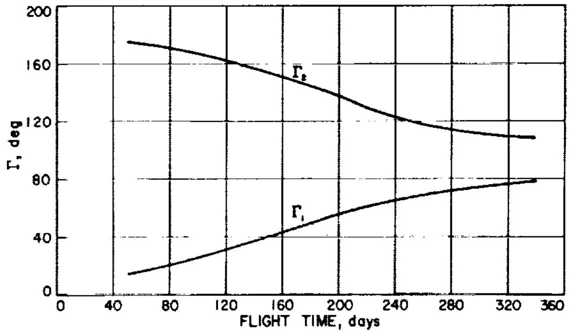


FIGURE 22-5.—Optimum Γ_1 and Γ_2 versus flight time.

directions, fixed relative to the radius vector, the increase in J for rendezvous trajectories departing near the optimum launch date is only 1 or 2%. Furthermore, the use of three or more allowed thrust directions gains very little in performance; the use of only one thrust direction for rendezvous trajectories is generally extremely inefficient, and in many cases the mission cannot be accomplished. For flyby missions, the use of only one thrust direction just slightly degrades the vehicle performance; this is because the variation of the optimal direction of thrust for typical flyby missions is much less radical than in rendezvous missions. From the computational standpoint, the variable direction thrust program is more convenient, since in resolving the two point boundary problem it is not necessary in this program to satisfy the conditions of Eq. (41).

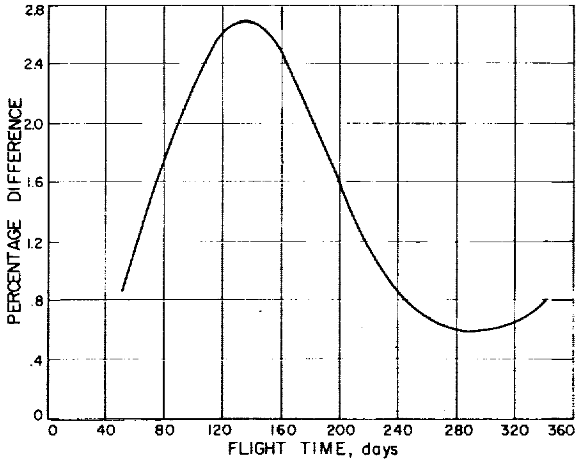


FIGURE 22-6.—Percentage increase in J with constant attitude program versus flight time.

A CURRICULUM FOR OPTIMIZATION THEORY

In view of the purpose of this NASA-University Conference and the need for better trained people of high caliber in the fields of systems optimization and optimal control theory, it seems appropriate to include a suggested outline for a year's course in this field. Unfortunately, all too few universities have developed strong curricula in these fields, which at this time are undergoing a dynamic and extensive growth. The course outlined below seems suitable for the senior or first-year-graduate levels. A list of texts which provide a considerable source of material is also included.

Course Outline for Optimization Theory

I. *Background Preliminaries*

Continuity considerations, differentiation, theory of maxima and minima, method of undetermined Lagrange multipliers, differentiation of integrals.

II. *Introduction to the Calculus of Variations*

The brachistochrone, minimum area of revolution, geodesics, isoperimetric problems.

III. *The Necessary Conditions for an Extremal*

Variational notation, basic lemmas, classical derivation of the Euler-Lagrange equations, Du Bois-Reymond's derivation, first integrals of the Euler-

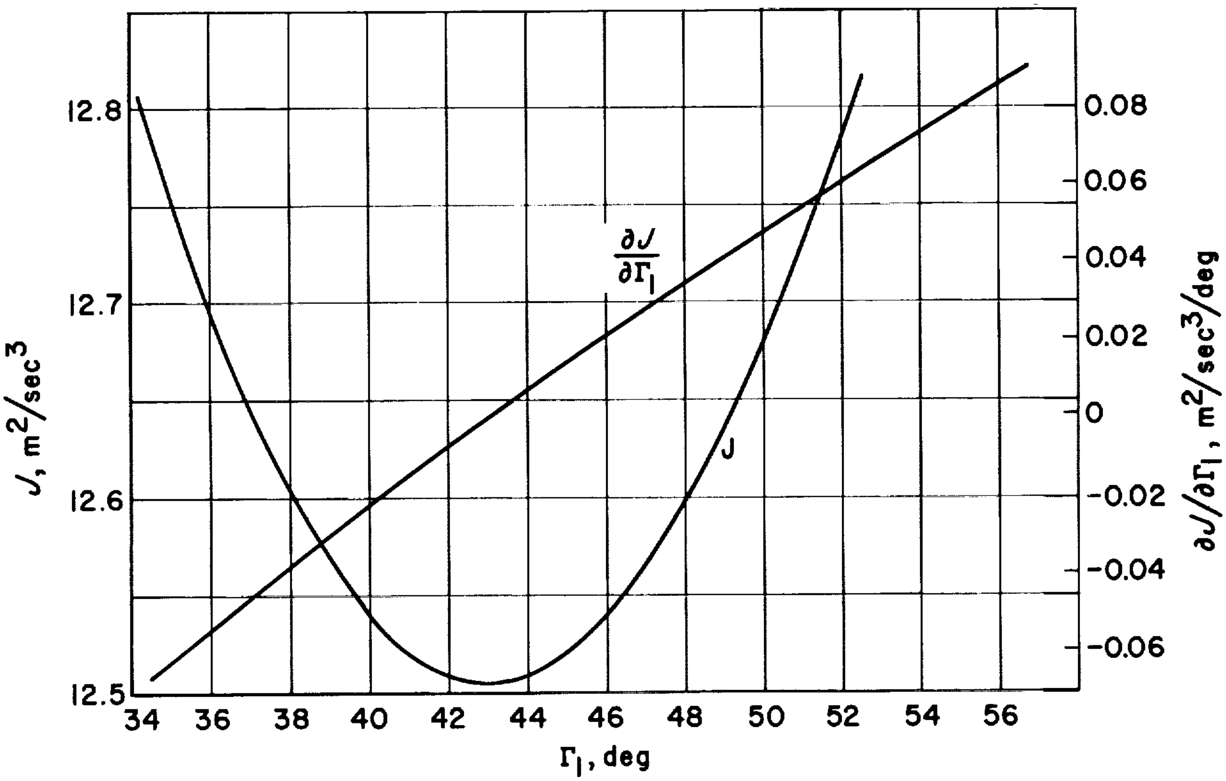


FIGURE 22-7.—J, $\partial J/\partial \Gamma_1$ versus Γ_1 for 160-day flight time.

Lagrange equations, Weierstrass-Erdman corner conditions.

IV. *Generalizations*

Multivariable analysis, higher derivatives in the integrand.

V. *Boundary Conditions*

Fixed end points, natural boundary conditions, end points on prescribed surfaces, transversality conditions.

VI. *Variational Problems with Accessory Conditions*

Lagrange multipliers, finite accessory conditions, differential accessory conditions, isoperimetric problems.

VII. *The Bolza, Lagrange, and Mayer Formulations of the Calculus of Variations*

The problems of Bolza and Mayer, the equivalence of the problems of Bolza, Lagrange, and Mayer. The multiplier rule.

VIII. *Additional Conditions for an Extremal*

Legendre's condition, Weierstrass E-function, Jacobi's condition.

IX. *Variational Problems with Inequality Constraints*

Bounded control variables, bounded state variables, Pontryagin maximum principle.

X. *Calculus of Variations and the Differential Equations of Mathematical Physics*

Hamilton's principle and Lagrange's equations of motion, Hamilton's equations, canonical transformations, Hamilton-Jacobi differential equation, Schrödinger equation, Fermat's principle, vibration problems, Sturm-Liouville problem, Rayleigh-Ritz method.

XI. *Methods of Numerical Solution of Variation Problems*

The two-point boundary value problem, methods of steepest descents, dynamic programming, indirect methods: search methods, adjoint variable techniques.

XII. *Applications to Space Flight Optimization*

Optimal rocket trajectory analysis with impulsive and continuous thrust propulsion systems, the effect of constraints on payload optimization, system parameter optimization, guidance theory along extremal paths.

Texts

BELLMAN, R., *Applied Dynamic Programming*, Princeton University Press, Princeton, 1962.

BLISS, G., *Lecture on the Calculus of Variations*, The University of Chicago Press, Chicago, 1946.

BOLZA, O., *Lectures on the Calculus of Variation*, G. E. Stechert and Co., New York, 1946.

COURANT, R., *Methods of Mathematical Physics*, Vol. I, Interscience Publishers, Inc., New York, 1953.

HILDEBRAND, F., *Methods of Applied Mathematics*, Prentice-Hall, New York, 1952.

LEITMANN, G. (ed.), *Optimization Techniques*, Academic Press, New York, 1962.

PONTRYAGIN, L. S., et al, *The Mathematical Theory of Optimal Processes*, Wiley, Interscience Division, New York, 1962.

WEINSTOCK, R., *Calculus of Variations*, McGraw Hill, New York, 1952.

REFERENCES

1. BOLZA, O., *Lectures on the Calculus of Variations*, G. E. Stechert and Co., New York, 1946.
2. BLISS, G. A., *Lectures on the Calculus of Variations*, The University of Chicago Press, Chicago, 1946.
3. VALENTINE, F. A., "The Problem of Lagrange with Differential Inequalities as Added Side Conditions", Dissertation, Department of Mathematics, University of Chicago, Chicago, 1937.
4. CICALA, P., *An Engineering Approach to the Calculus of Variations*, Levrotto and Bella, Torino, 1957.
5. HESTENES, M. R., "A General Problem in the Calculus of Variations with Applications to Paths of Least Time", The RAND Corporation, Rpt. No. RM–100, Santa Monica, 1950.
6. PONTRYAGIN, L. S., et al., *The Mathematical Theory of Optimal Processes*, John Wiley and Sons, Inc., Interscience Division, New York, 1962.
7. BLISS, G. A., *Mathematics for Exterior Ballistics*, John Wiley and Sons, New York, 1953.
8. MELBOURNE, W. G., RICHARDSON, D. E., and SAUER, C. G., JR., *Interplanetary Trajectory Optimization with Power-Limited Vehicles*, Technical Report No. 32–173, Jet Propulsion Laboratory, Pasadena, 1961.
9. COURANT, R., *Variational Methods for the Solution of Problems of Equilibrium and Vibrations*, Bull. Am. Math. Soc., 49, 1943, pp. 1–23.
10. COURANT, R., and HILBERT, D., *Methods of Mathematical Physics*, Interscience Publishers, New York, 1953.
11. KELLEY, H. J., *Gradient Theory of Optimal Flight Paths*, J. Am. Rocket Soc., Vol. 30, October 1960, pp. 947–954.
12. BRYSON, A. E., CARROLL, F. J., MIKAMI, K., and DENHAM, W. F., "Determination of the Lift or Drag Program that Minimizes Re-entry Heating with Acceleration or Range Constraints Using a Steepest Descent Computation Procedure," Paper presented at IAS 29th Annual Meeting, New York, 23–25, 1961.
13. BELLMAN, R., *Dynamic Programming*, Princeton Univ. Press, Princeton, 1957.
14. BELLMAN, R., *Adaptive Control Processes: A Guided Tour*, Princeton Univ. Press, Princeton, 1961.
15. BELLMAN, R., and DREYFUS, S. E., *Applied Dynamic Programming*, Princeton Univ. Press, Princeton, 1962.
16. DREYFUS, S., *Variational Problems with Inequality Constraints*, J. Math. Anal. and Appl., Vol. 4, 1962, pp. 297–308.
17. LAWDEN, D. F., *Advances in Space Science*, ed. F. I. Ordway III, Academic Press, New York, 1959, Vol 1, Chap. 1.

18. LEITMANN, G., *On a Class of Variational Problems in Rocket Flight,* J. Aero-Space Sciences, Vol. 26, 1959, p. 586.
19. MIELE, A., *An Extension of the Theory of the Optimum Burning Program for the Level Flight of a Rocket-Powered Aircraft,* J. Aero. Sci., Vol. 24, 1957, p. 874.
20. BREAKWELL, J. V., *The Optimization of Trajectories,* J. Soc. Indust. Appl. Math., Vol. 7, No. 2, June 1959, pp. 215–247.
21. IRVING, J. H., *Space Technology,* ed. H. S. Seifert, Wiley and Sons, New York, 1959, Chap. 10.
22. MELBOURNE, W. G., and SAUER, C. G., JR., *Optimum Thrust Programs for Power-Limited Propulsion Systems,* Technical Report No. 32–118, Jet Propulsion Laboratory, Pasadena, June 15, 1961, also Astronautica Acta, in press.
23. MELBOURNE, W. G., and SAUER, C. G., JR., *Payload Optimization for Power-Limited Vehicles,* Technical Report No. 32–250, Jet Propulsion Laboratory, Pasadena, April 9, 1962, also American Rocket Society Progress Series, Vol. 9, *Electric Propulsion Development,* in press.